arXiv:math/0603228v1 [math.ST] 9 Mar 2006

NONPARAMETRIC BAYESIAN CLASSIFICATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Marc A. Coram
August, 2002

© Copyright by Marc A. Coram 2002
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Persi Diaconis
(Principal Adviser)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jerome H. Friedman

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Bradley Efron

Approved for the University Committee on Graduate Studies:

Abstract

A Bayesian approach to the classification problem is proposed in which random partitions play a central role. It is argued that the partitioning approach has the capacity to take advantage of a variety of large-scale spatial structures, if they are present in the unknown regression function f_0. An idealized one-dimensional problem is considered in detail. The proposed nonparametric prior uses random split points to partition the unit interval into a random number of pieces. This prior is found to provide a consistent estimate of the regression function in the $\mathcal{L}^p$ topology, for any $1 \leq p < \infty$, and for arbitrary measurable $f_0 : [0,1] \to [0,1]$. A Markov chain Monte Carlo (MCMC) implementation is outlined and analyzed. Simulation experiments are conducted to show that the proposed estimate compares favorably with a variety of conventional estimators. A striking resemblance between the posterior mean estimate and the bagged CART estimate is noted and discussed. For higher dimensions, a generalized prior is introduced which employs a random Voronoi partition of the covariate-space. The resulting estimate displays promise on a two-dimensional problem, and extends with a minimum of additional computational effort to arbitrary metric spaces.

Acknowledgements

I thank my parents, who laid the foundations of my character and have been an unfaltering source of love and encouragement. I also thank my advisor, Persi Diaconis, whose friendship, support, and insight have been vital to successfully enduring the Ph.D. process. I also wish to thank the other members of my examining committee: Jerome Friedman, Bradley Efron, Susan Holmes, David Siegmund, and Hans Andersen, for their professionalism and patience. Finally, I thank Hua Tang for putting up with my quirks all these years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The binary classification problem is perhaps the simplest regression problem, but it continues to pose fresh challenges. In the binary classification problem, we are given a list of n pairs $Z_i = (X_i, Y_i)$ each pair drawn independently from an unknown probability measure F. The X's play the role of covariate or "predictor" and lie in some abstract space $\mathcal{X}$, while the Y's are interpreted as a class label and are either 0 or 1. Our goal is to estimate certain functionals of F. Specifically, in the binary regression problem, we are interested in estimating the regression function $f : \mathcal{X} \mapsto [0, 1]$. The value of f at a given point $x \in \mathcal{X}$ is the conditional probability that $Y = 1$ given that $X = x$. In this way we model the joint distribution F by saying that to draw an (X, Y) pair from F, first draw a covariate $X = x$ from the marginal distribution of X denoted by μ. Then "flip" an $f(x)$ coin to determine the value of Y.

In the classification problem, we are concerned with being able to predict future Y values. The standard formalization of this task is that we wish to choose the "decision rule" that will minimize the expected loss incurred; this reduces to the problem of estimating the set $\{x \in \mathcal{X} : f(x) > c\}$, for some c that depends upon the loss (for simplicity, ignore the possibility that c depends on x). There are a great many ways to proceed on each of these problems, as demonstrated by the vast literature on these subjects. Some references are

given in section 2.3.

In this thesis, I propose a nonparametric Bayesian approach to the binary classification and regression problems. Specifically, to derive an estimator, I regard F itself as random. For simplicity, I regard the marginal distribution of X, as known. In this case, putting a prior distribution on F amounts to putting a prior on f. More generally, one can also put a prior on functions m and suppose that $\mu(dx) = m(x)\mu_0(dx)$. Some sort of mild restriction, like this one that all μ share some dominating measure μ_0, is useful to avoid technical problems in defining the conditional distribution of F given the data.

Let π denote a prior distribution on F, or, more precisely, on (f, m) pairs. Extend π to a joint distribution on $F = (f, m)$ and the infinite data sequence $(Z_1, Z_2, \dots)$ which, conditionally on F, is drawn independently and identically distributed (*iid*) from F. Formally, the posterior is the measure $\pi_n(dF) := \pi(dF|Z_1 = z_1, \dots, Z_n = z_n)$. In practice Markov chain Monte Carlo procedures can be used to generate a sample from the posterior.

The posterior mean of f is an important summary of the posterior: its value minimizes the posterior risk under an $\mathcal{L}^2$ loss. Let $\widehat{f}$ denote the posterior mean: $\widehat{f}(x) = \int f(x)\pi_n(df)$. Another important summary of the posterior is the classification rule which minimizes posterior misclassification loss. If asked to predict the most likely value of the Y's corresponding to $X_{n+1}, \dots, X_{n+n'}$ all at once, the decision that would minimize the posterior-expected 0-1-loss is simply $\delta_i = \mathbf{1}_{\widehat{f}(X_i) > \frac{1}{2}}$. Interestingly, though, if asked sequentially instead of all at once, it is necessary to update $\widehat{f}$ with each new data point before deciding.

Taking this Bayesian approach assures us that the resulting estimators will have a clear subjective interpretation. In addition, if the prior π is carefully chosen, the resulting estimators, chosen indirectly through this Bayesian framework, may have frequentist advantages over the estimators that might otherwise be proposed. For example, interesting kinds of shrinkage and averaging

occur automatically within this framework. Subsequent chapters assess the frequentist performance of these Bayesian estimates by simulation experiments (chapter 4) and theoretically (chapter 5).

1.1 An Example

To get started, let us consider the specific case in which $\mathcal{X} = [0, 1]$ and the sampling distribution of the X_i, μ, is known to be the $U(0, 1)$ distribution. Further, let us consider a specific $f = f_0$ which is complicated enough that its estimation should not be too easy for any of the standard methods; it is piecewise continuous with two constant regions and a smooth transition region. As shown in Figure 1.1.a, $f_0(x)$ is chosen as:

$$f_0(x) = \begin{cases} 0 \leq x < \frac{1}{6} & 0.6 \\ \frac{1}{6} \leq x \leq \frac{1}{2} & 0.4 \\ \frac{1}{2} < x \leq 1 & \frac{\phi_\sigma(x-\frac{1}{2})}{\phi_\sigma(x-\frac{1}{2})+\phi_\sigma(x-1)} \end{cases}$$

where ϕ_σ is the density of a normal with mean 0 and standard deviation $\sigma = 0.25$.

The two histograms in Figure 1.1.a summarize a simulated data set of 1024 data points that was drawn from this model. The green histogram is a histogram of the heads. The red histogram is a histogram of the tails; it is drawn upside down to facilitate comparison with the green histogram. To more easily interpret this display, notice that if we take the sum of the corresponding green and red bins at each point, we recover a histogram of the marginal distribution, which is uniform. Furthermore, the ratio of the height of a green bin to the corresponding red bin represents the empirical odds of a head in that bin. Near $x = 0.5$, for example, notice the sharp transition from nearly equal green and red bins (on the left) to much longer green than red bins (on the right). The performance of this posterior mean estimator is compared with a variety

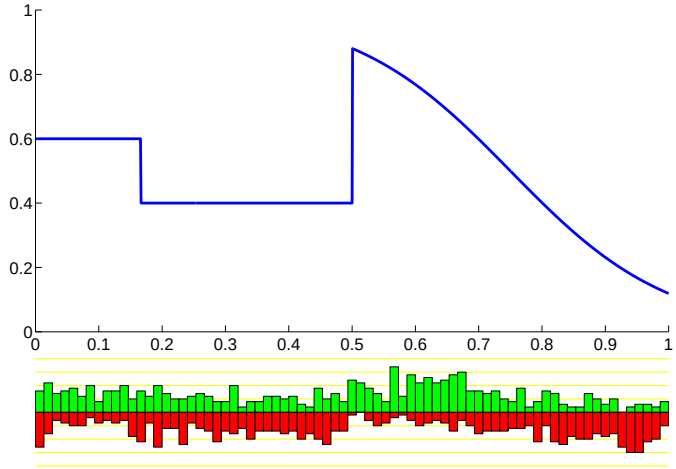

Figure 1.1.a: An Example: $f_0(x)$

of more conventional methods in chapter 4. This example is used to illustrate the present approach in the rest of this introduction.

1.2 A Nonparametric Prior on Regression Functions

To specify a prior π over functions $f : [0,1] \mapsto [0,1]$, I explain how to choose f, at random from it. This completely specifies a prior on the probability distribution F, since I consider the marginal distribution μ to be known. The prior on f will concentrate on locally-constant step functions. To choose a step function at random, first, choose K, the number of locally constant intervals, where:

$$P(K = k) = (1 - \alpha)\alpha^{k-1} \quad \text{for } k = 1 \ldots \infty$$

That is, K is *Geometric* with parameter $1 - \alpha$. Ultimately, the choice of α must be specified by the user. For the examples in this thesis, I have used $\alpha = \frac{1}{2}$ unless otherwise noted. This choice seems to perform well. For further discussion of how to choose a prior on K which results in provably consistent

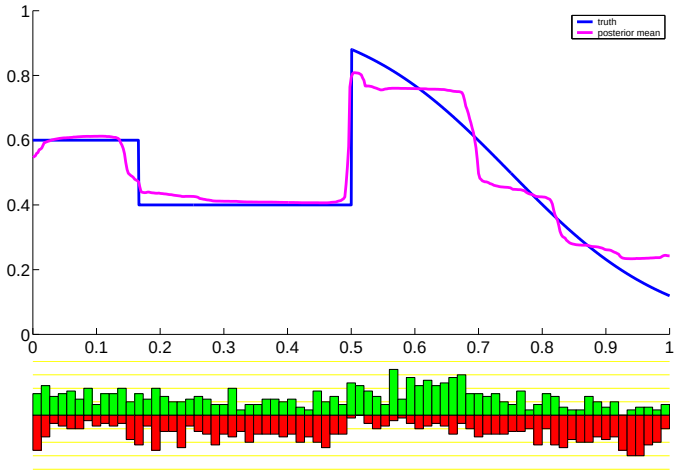

Figure 1.3.a:

estimators, see chapter 6.

Now, conditional on $K = k$, choose $V_1, V_2, \ldots, V_{k-1}$ *iid* from $U(0,1)$. Let $V_{(i)}$ (for $i = 1 \ldots k-1$) be the $i'th$ ordered value of the V's. This produces k intervals:

$$I_1 = \{0 \leq x < V_{(1)}\}, I_2 = \{V_{(1)} \leq x < V_{(2)}\}, \ldots, I_k = \{V_{(k-1)} \leq x \leq 1\}$$

If $k = 1$, simply take $I_1 = [0,1]$. Finally, conditional on $K = k$, choose k values S_i (for $i = 1 \ldots k$) *iid* from $U(0,1)$. This generates the random function f:

$$f(x) = \sum_{i=1}^{k} S_i \mathbf{1}_{x \in I_i}$$

1.3 Sample Results

Conditioning this prior on the data described in the earlier example section results in a posterior distribution on regression functions f. Applying the techniques in chapter 3 to sample from the posterior results in a long list of

sampled functions. Although the prior specified a $Geometric(\frac{1}{2})$ prior distribution on the number of locally-constant pieces K in each function, when drawn from the posterior, functions typically have at least 5 pieces. Taking the average of these functions (at every x) gives an estimate of the posterior mean; this is illustrated in Figure 1.3.a. Further discussion of results like this can be found in chapter 4.

1.4 The Partitioning Approach

Step functions, such as the functions f that the prior π concentrates on, have the advantage of great mathematical simplicity; but, if specified in sufficient detail, they can approximate general functions. For a multi-dimensional regression function, it is natural to generalize this idea by considering some partition of the covariate space $\mathcal{X}$ into a number of pieces and constructing a function that takes a different value on each piece. A procedure that uses a wide variety of geometric shapes to partition the space might be able to find a partition with the right structure to approximate the unknown regression function. Ideally, the partition would be no more complex than necessary to achieve a good approximation. If the unknown regression function has certain global features that the chosen partition can be adapted to, this idea becomes very powerful. Instead of merely "borrowing strength" locally, like ordinary smoothing estimators do, a partition-based estimator borrows strength across the whole range of a partition element. As a simple example, if the true regression function does not depend on one of the covariates, the partition elements do not need to break up space along this dimension at all; this results in larger partition elements and more efficient estimation of the success probability on each of the pieces. Similarly, if the regression function is almost flat in some large chunk of space, then this whole region can become a single element. If the level-sets of the regression function have smooth boundaries, perhaps the partition elements can be chosen to follow these contours. Finally, since the best partition

for the unknown function is unknown, it makes sense to average together the approximations found by a variety of partitions that make the data likely. This is exactly what the posterior mean estimate will do automatically. Chapter 7 shows one way in which this idea can be applied using Voronoi partitions. For this partition, a prior distributes seed-points in the covariate space; each seed corresponds to a partition element, the one that consists of all points closer to that seed than any other seed. By placing the seeds appropriately, the partition can be fine or coarse as needed. The boundary between partition elements is itself a hyperplane whose position and orientation can be controlled through the placement of the seeds.

Other authors have recently considered similar priors with good success. Their work is discussed in section 2.2. To the best of my knowledge, this thesis presents the first theoretical examination of consistency issues for priors of this sort (c.f. chapter 5, with discussion in chapter 6). Additionally, I present a detailed assessment of the empirical performance of my methods on certain novel simulation experiments. The comparison with bagged CART regression trees in chapter 4 is especially interesting.

1.5 Outline

Chapter 2 gives a literature review. Chapter 3 shows how to (approximately) compute samples from the posterior and the posterior mean. Chapter 4 trys out the method on examples and carefully compares its performance with that of a variety of existing methods. Chapter 5 gives sufficient conditions on the prior under which it provides universally consistent estimates of f. Chapter 7 describes a different prior which extends these ideas to general metric spaces by employing random Voronoi partitions. Certain modifications are explained that make the proposal more practical and its performance on an example is shown. Finally, the afterword gives a philosophical argument that advocates the use of Kolmogorov-complexity in future statistical thinking.

Chapter 2

Literature

This chapter reviews and discusses the literature on three subjects. The first section reviews some theoretical results concerning the frequentist performance of Bayesian procedures. The second section gives a survey of some of the work done by authors on related Bayesian efforts. The final section briefly surveys some salient examples of alternative approaches to the classification problem.

2.1 Theoretical Results

The frequentist performance of Bayesian methods is of fundamental interest in statistics. Given a large sample from a smooth, finite-dimensional statistical model, the situation is quite well understood. The Bernstein-von Mises theorem [49, 33] shows that the Bayes estimate and the maximum likelihood estimate will be close. Furthermore, the posterior distribution of the parameter vector around the posterior mean is close to the distribution of the maximum likelihood estimate around the truth: both are asymptotically normal with mean 0 and the same covariance matrix. Unfortunately, though, in more general circumstances, such as those needed for this work, the situation can be much more complex. In particular, the basic model is based on an infinite

hierarchy of finite dimensional models. Moreover, even for a given finite dimensional submodel, the dependency of the likelihood function on the parameter is not smooth; the functions are allowed to take jumps. Consequently, a more general theory is needed.

This section reviews some of the literature on this subject with a focus on results that address the question of consistency: i.e. as the number of data points tends to infinity, will the Bayesian estimate converge to the true value (in some suitable sense) almost surely (resp. in probability)? The literature contains a number of useful and quite flexible positive results, but also a variety of interesting negative examples showing that the regularity conditions under which the theorems hold are not to be taken lightly. A good introduction to these issues is by Diaconis and Freedman [23]. Throughout this section, the reader may envision a family $\mathbf{P}_\theta(dx)$, a prior $\pi(d\theta)$, and posterior $\pi(\theta|x_1, \ldots, x_n)$, where the x_i are drawn *iid* from $\mathbf{P}_{\theta_0}(dx)$. Consistency means that the posterior concentrates at θ_0 for large samples.

Doob [29] established a fundamental result under minimal regularity assumptions using a martingale convergence argument. Roughly speaking, the result states that if consistent estimators exist at all, then a Bayes procedure will provide an almost surely consistent estimate of the true parameter θ under sampling from the θ distribution for any θ in some set B which has prior probability of 1. Notice, though, that this does not specify if consistency will obtain at any *particular* point of interest θ_0, unless θ_0 happens to be a point-mass of the prior, or unless it possible to determine B by some more detailed line of argumentation.

Freedman [32] considered the case in which the observations are discrete. If the set of possible observations is finite, the posterior is consistent exactly for parameter values in the topological support of the prior. The countably infinite case is more complex. He constructs a class of examples showing that it is possible to construct a prior which assigns positive mass to every (weak star) neighborhood of the true parameter value, but for which the posterior converges

to a point mass at some other (chosen) parameter value. Furthermore, he finds a prior which assigns positive prior mass to every (weak star) open set of parameters, but for which the posterior is consistent only at a set of parameters of the first category. The reader should note that this prior did not assign mass to all entropy-neighborhoods. This sort of subtle distinction can make all of the difference and explains the necessity of some such assumption in the following consistency theorems. He introduces the "tail-free" priors for the the countably-infinite case and demonstrates that these are always consistent.

Lorraine Schwartz [61] explored the question of consistency in a very general setting. She extended Doob's result to a broad class of loss functions [61, lemma 4.2]. She also found sufficient conditions for the posterior to be consistent under *iid* sampling. These conditions, she says, are "of an essentially weaker nature" than the conditions established for the consistency of maximum likelihood estimators. Nevertheless, she constructs an example where the maximum likelihood estimate is consistent and the estimates based on certain priors are not. The example ([61, example 3]) involves a simple parametric family of densities which satisfies Wald's conditions, thereby guaranteeing that the maximum likelihood estimate will be consistent, but for which the posterior can be inconsistent. The consistency of the posterior in this case, is found to depend critically on the amount of mass that the prior ascribes to small neighborhoods of the true parameter value; if this mass shrinks too quickly, the prior "ignores" the data. One clever aspect of her construction is the way the densities are parametrized. Parameter values close to the target value θ_0 correspond to densities that are close to the θ_0-density in an $\mathcal{L}^1$ sense, but which are farther and farther away in Kullback-Leibler discrepancy. In fact, there is only one point in parameter space (the true parameter) that has Kullback-Leibler discrepancy from the truth smaller than ϵ, for ϵ sufficiently small.

Schwartz then shows that the posterior will be consistent under *iid* sampling under two basic conditions. First, the prior should have positive mass on Kullback-Leibler neighborhoods of the true parameter (defined in section 5.1

of this thesis), and second, the model class should not be too rich; specifically, she requires that uniformly consistent tests of the hypothesis that $\theta = \theta_0$ against the alternative that θ lies outside a given (open) neighborhood of θ_0 exist.

It is not always obvious how to verify the later property directly. Modern authors have employed entropy-type bounds to guarantee their existence. Ghoshal, Ghosh, and van der Vaart [40] state a theorem ([40, theorem 7.3]) which proves that the posterior converges at a certain rate if certain uniform tests exist (and the prior mass is suitably distributed) and go on to find a variety of entropy-type conditions that suffice to be able to construct the necessary tests. Shen and Wasserman [62] show related results, requiring slightly different conditions on how mass needs to be allocated–they do not a make a connection with testing. Barron, Schervish, and Wasserman [2] find sufficient conditions for the posterior to be consistent; their results are reviewed and then used in chapter 5.

It should be noted that these various conditions for consistency are not necessary, but merely sufficient. Nevertheless, it is important to treat this subject with care because of the variety of examples for which consistency fails.

Barron, Schervish, and Wasserman also give an interesting example where consistency fails. In this example, they show that the prior puts too much mass on a very rich class of models that will be able to match any spurious structure that the data might have by chance, overwhelming the true parameter. Furthermore, lest the reader get the wrong idea, inconsistency does not only occur in artificial examples. A series of "natural" yet still inconsistent estimators for the symmetric location problem are discussed by Diaconis and Freedman [23]. In addition, the binary regression example explained in the next section has a natural motivation based on conditional exchangability.

2.2 Related Bayesian Work

The following subsections contain a review of work by other authors that is closely related to this thesis. It is followed my a brief synopsis of the contributions that this thesis makes to the literature.

2.2.1 A Dyadic Prior for Binary Regression

The most relevant examples for the work of this thesis are the nonparametric binary regression examples of Diaconis and Freedman [24, 25]. They use a different prior; call it π_{DF}, a hierarchical, dyadic prior on f. To describe π_{DF}, let A_k be the set of intervals which result from partitioning the unit interval into 2^k equal pieces. Let $\mathcal{F}_k$ be the subset of functions which are constant on all intervals $a \in A_k$. Finally, fix a prior distribution κ on the non-negative integers. Assume, for simplicity, that $\kappa(k) > 0$ for all k. To draw f from π_{DF}, draw K from κ and then, conditional on hierarchy-level $K = k$, draw f uniformly at random from $\mathcal{F}_k$. In effect then, at level k one draws 2^k independent $U(0,1)$ random variables to describe the success probability on each of the 2^k pieces.

They show that for any κ and any f_0 (except possibly for $f_0 \equiv \frac{1}{2}$), the posterior estimates are consistent (in the sense that any $\mathcal{L}^1$ neighborhood of f_0 has posterior probability tending to 1 a.s.). Remarkably, however, for $f_0 \equiv \frac{1}{2}$, the posterior can be an inconsistent estimate if the tail of κ is sufficiently heavy. Specifically, let $\lambda_k = -\log(\kappa(K \geq k))/k$. Then if $\limsup \lambda_k > \lambda_{crit} = 2^{-\frac{1}{4}} \approx 0.841$, the posterior is inconsistent at $f_0 \equiv \frac{1}{2}$. On the other hand, if $\limsup \lambda_k < \lambda_{crit}$, the posterior is consistent for any f_0. To put this in perspective, for $\kappa(k) = (1-\beta)\beta^k$ (a shifted *Geometric*$(1-\beta)$ prior), $\limsup \lambda_k = -\log(\beta)$. The critical value for β is $\exp(-\lambda_{crit}) \approx 0.431$; for larger β (longer tails) inconsistency will occur (but only for $f_0 \equiv \frac{1}{2}$).

This result is substantially stronger than the result I have obtained for my prior π. In particular, applying the same (general) method of proof that I employed to prove consistency for π to π_{DF} yields only the result that π_{DF} is

consistent if the tails of κ drop off at least as fast as those of a Poisson. (Recall, that at level k, π only divides $[0,1]$ into k intervals, but π_{DF} divides it into 2^k.) Their method of proof is direct: using Bernstein's inequality, Poissonization, and special features of the prior. My method of proof is indirect; it uses general results that employ entropy-type bounds.

There are striking similarities between π and π_{DF}. In fact, π is equivalent to a suitably randomized π_{DF}. To achieve this, it is not enough to simply randomize the dyadic split points. Instead, recall that π_{DF} has an alternative interpretation in terms of binary sequences. At hierarchy-level k, π_{DF} is uniform over $\mathcal{F}_k$. This corresponds to independently assigning uniform success probabilities to each binary sequence of length k. Here is an alternative way to draw f from π. Draw g from π_{DF} and interpret g as function on binary sequences of length k (k depends on g). Let V_i $(i = 1, \ldots, k)$ be *iid* $U(0,1)$ random variables. To any point $u \in [0,1]$ associate the binary random variables $\eta_i(u) = \mathbf{1}(u \leq V_i)$ $(i = 1, \ldots, k)$. Define f via $f(u) = g((\eta_1(u), \ldots, \eta_k(u)))$. Note that only a small fraction of possible binary sequences are realized in this manner (at level k (which ranges from 0 to ∞ under π_{DF}), $k+1$ sequences out of the full set of 2^k possible sequences are achieved).

2.2.2 Bayesian CART

Two other closely related priors can be described as Bayesian versions of the CART algorithm. This was pursued by Chipman, George, and McCulloch, whose prior closely parallels the choices made in the original CART algorithm [6, 7, 8, 9]. Here is a description of their prior when the covariate space is $\mathcal{R}^p$. Their prior starts with a root node (which represents the whole space); this node is then recursively partitioned in a random way. For each node, randomly choose whether to split it or not, then choose a coordinate to split on, then choosing a split point (i.e. the cutoff value) randomly *from among the midpoints between the ordered values of this coordinate*; finally each

leaf node is given an independent regression value. The details of how these decisions are made differ in their particulars from the ones that I described in the introduction. In early work, these authors observed that using MCMC to sample from the posterior of this prior provides a rudimentary (global) search procedure, which has certain (apparent) advantages over the *greedy* search procedure commonly implemented in CART-type algorithms. In later work, they examined and computed the (approximate) posterior mean (working primarily on the least-squares white-noise regression problem) and found that it had good performance. They also considered extended priors that modeled the regression values as additively (not independently) generated [9].

Denison, Mallick, and Smith, independently considered another version of Bayesian CART [18, 17, 19]. For one-dimensional problems they propose using random splines (the prior I use is essentially a special case of this prior). They consider some of the regression examples that are standard in the wavelet literature and show that their spline methods perform equally well. Additionally, they propose a Bayesian version of Friedman's MARS which puts a prior on functions that are constructed by adding together random spline-type ridge functions. Denison, Adams, Holmes, and Hand discuss the usefulness of random partitions in this paper [15].

Very recently, Denison, Holmes, Mallick, and Smith have written a book [16] which surveys some related Bayesian regression schemes, including a Bayesian method for (multiple class) classification using Voronoi partitions that is very closely related (albeit independent of) the work that I present in chapter 7. The book also discusses Bayesian wavelet methods, and an interesting Bayesian nearest-neighbor prior. As a default prior, they recommend assuming that every model in a "single dimension" is equally likely, and each dimension is equally probable, *a priori*. This "flat prior," they claim, should serve perfectly well because of the, "natural tendency" for the marginalized likelihood to penalize complex models:

> On the face of it, we might be concerned that the flexible modeling strategy we advocate might be prone to overfitting the data by adding too many basis functions. Indeed, many papers found in the literature advocate explicit priors on the model space that penalize the dimension of the model. However, throughout this book we argue that such a measure is unnecessary. The Bayesian framework contains a natural penalty against over [sic] complex models, sometimes called *Occam's razor*, which essentially states that a simpler theory is to be favoured over a more complex one, all other things being equal.

There is no consideration given to the possibility that this might give rise to inconsistent estimates (e.g. as in the Diaconis and Freedman non-parametric regression example explained earlier); indeed there are few theoretical considerations at all in the book. Their explanation of why the Markov chain techniques that they develop should actually give meaningful samples from the posterior appeals to Green's reversible jump [41]. The explanation given is vague and ultimately they decide to avoid the issue and appeal to the fact that their chains are discrete. The chains in chapter 3 of this thesis involve a continuous state space and do not simply avoid this issue by discretizing the continuous modeling space as these authors seem to do.

Overall, the book emphasizes main ideas, algorithms, and results. It seems that for every existing regression technique, they want to demonstrate that they can make a "Bayesian" version of it too. The book does not emphasize subjectivism, but rather adopts an "$\mathcal{M}_{open}$" perspective to Bayesian modeling: "we never believe that the true model lies in the set of possible models." The book does do a good job of supplying default priors for a wide variety of possible parametric models. Similarly, Denison's thesis [20] emphasizes the wide variety of problems to which Bayesian partitioning methods of this sort can be applied.

2.2.3 Poisson Rate estimates using Random Partitions

Green [41], and Scargle [58] develop priors on piecewise constant functions on the real line and $\mathbb{R}^d$ using Voronoi cells. Their priors are quite similar to the ones developed in this thesis, but are intended to address the problem of estimating the rate function of a *Poisson* process. In principle, one could apply their techniques to the problem of binary regression by generating an estimate of the rate function of the "heads" process and the "tails" process separately and then combining the results. I do not think that this has been tried and it seems substantially less "natural."

Green applies his method to a coal mining dataset and a synthetic two-dimensional example. For these example, Green assumes that an individual cell's rate-parameter is drawn independently from a $\Gamma(\alpha, \beta)$ prior. For the one-dimensional case he advocates a prior which "probabilistically" spaces out the change-point locations; specifically, if there are j change-points, the ordered locations of the change-points are distributed like the even order statistics of $2j + 1$ independent uniform values. He argues that this is good because it prevents small change-point intervals from entering into the posterior. For the two-dimensional example, the generating points of the Voronoi partition are drawn independently and uniformly. Green's methods are given, in part, as examples of his "reversible jump" MCMC technique. This technique has become an accepted part of MCMC practice, but is not accepted by all experts in MCMC theory because it does not lay down in a straightforward "theorem-proof" manner the necessary conditions and consequent conclusions. For this reason, detailed verifications for the chains used in this thesis are given in chapter 3.

Scargle's work is applied to astronomical data; he concentrates on the problem of finding the mode of the posterior, rather than the posterior mean. Fortunately, he and coworkers have developed a way of computing this mode in the one-dimensional case exactly and efficiently using a dynamic programming

approach [59]. Instead of giving each cell an independent value, Scargle gives each cell a (logical) "color" and then associates each unique color with an independent rate-parameter. This allows him to use a fine partition and then group "chunks" back together into more complicated shapes. The way he forms this partition is also different; in particular his "prior" is data dependent, but not quite in the way of the "prior" that I consider in chapter 7. Rather, the data is used once and for all to generate the fine Voronoi partition of space that results from using all of the data points as generators. These cells are then "clumped" (i.e. given a logical color) and the clumps are given an independent rate parameter.

2.2.4 Bayesian "Image" Analysis

Møller and Skare [53] apply their work to reservoir modeling and connect their work to efforts in Bayesian image analysis (including Markov random fields). They use a random Voronoi partition of the data and assign each partition element a random color (in a way that depends only the colors of neighboring cells). They supply several further references to work in Bayesian image analysis which use Voronoi cells. From their perspective, to calculate their posterior they are simulating from a special "marked point" process. The generators of the Voronoi cells are regarded as point set that has been drawn from a homogeneous Poisson process of rate β on the unit cube. In the simplest case, the marks or "colors" of these points are just integers from 1 up to M that have been drawn independently. More generally, according to their prior, the conditional distribution of the coloring of cells given is an Ising or Potts model. The graphical structure of this model is determined by consideration of which Voronoi cells are neighbors, and the θ parameter is chosen to reflect their prior belief that neighboring cells tend to be of the same color. They consider two problems. The first is a simulation experiment in which a "true"

binary image is degraded with Gaussian noise. The second is a three dimensional reservoir problem based on real data. It is supposed that a certain three dimensional cube (the reservoir) consists of 4 different types of rock. The rock types are observed along seven vertical lines, representing the observations of rock that were made as seven wells were dug into the reservoir. In both problems, the true object to be recovered is itself a certain "coloring" of space (i.e. rather than a continuous regression function). For the MCMC computation of their posterior they apply the birth-death type Metropolis-Hastings algorithm for point processes, as studied by Geyer and Møller [38] and claim that their target distribution satisfies a local stability condition (see Geyer [37], Kendal and Møller [45], and Møller [52]) so that the MCMC is actually geometrically ergodic.

2.2.5 Polya Trees

Finally, Polya trees [48] and especially randomized Polya trees [55] deserve to be mentioned. The basic Polya tree puts a prior on distribution functions on the unit interval. The unit interval is divided recursively in a dyadic binary way and mass is allocated to each piece of the partition in a stagewise manner by first determining how much of the mass that is available will be on the left versus the right half and then continuing with such determinations layer by layer. Each of these assignments is ultimately determined by independent *Beta* random variable, whose parameters depend upon its location in the "tree." If a suitable choice of these parameters is made the result prior on distribution functions concentrates on distributions that are absolutely continuous with respect to Lebesgue measure. The essential advantage of Polya trees is that the posterior of Polya tree prior is easily and analytically computable, being itself another Polya tree. For randomized Polya trees, the partitioning scheme is independently "jittered" at random in a particular way [55]. A Hybrid MCMC can be employed to sample from the randomized Polya tree posterior which

uses a Gibbs step to take advantage of the ease with which the (internal) Polya tree posterior can be computed. Both methods can be extended (essentially by taking "direct products") to put a prior on distributions on the unit cube.

2.2.6 The Contributions of this Thesis

Reviewing the depth and breath of the literature reviewed above may leave the reader in doubt about the contributions of this thesis. After all the one-dimensional prior that I consider is essentially a special case of the univariate spline model and the idea of using Voronoi partitions is certainly not new, although effective Bayesian methods using them only started springing up fairly recently.

Still there is room for careful analysis. This thesis establishes that the posterior is consistent under suitable conditions on the prior and for any measurable regression function (see chapter 5 for details): an issue which none of the "Bayesian CART" or "Voronoi Partition" authors address at all. This thesis also gives an explicit Markov chain Monte Carlo algorithm (see section 3.4). Broadly speaking it is a fairly standard birth-death Markov chain as considered by Geyer and Moller [38], but the technicalities of the analysis seem to be somewhat different. This thesis proceeds to show in detail that it satisfies detailed balance by direct self-contained argumentation; further, the chain is shown to have an ergodicity property (see section 3.10). These considerations are often glossed over in modern writing.

On the more practical side, chapter 4 scrutinizes the behavior of the posterior mean estimate under a variety of carefully designed simulation experiments. These experiments both serve to analyze the posterior mean and to give insight into the relationship between Bayesian methods and their classical counterparts. See for example the discussion of CART and bagging in subsection 4.1.3.

2.3 Other Approaches

The literature on classification and regression methods is huge; the interested reader is urged to consult good modern books on the subject like *The Elements of Statistical Learning,* by Hastie, Tibshirani, and Friedman [43]. The following paragraphs outline some of the methods that have had the most impact upon the author.

In the statistics literature, classical approaches to the classification and binary regression problem include logistic regression, Fisher's discriminant analysis, and projection pursuit methods. Logistic regression specifies that the success probability regression function is such that its log-odds follows a linear model with a user specified basis (e.g. by using polynomial or spline functions of the covariate-data) and estimates the parameters by maximum likelihood. Model selection is commonly performed using classical methods to select a subset of the covariate variables. Fisher's discriminant analysis finds a hyperplane which "optimally" separates the two classes using a within versus between variance criterion. Projection pursuit seeks an interesting linear (or sometimes nonlinear) projection of the covariate-data onto a lower dimensional subspace (e.g. $\mathbb{R}$). Various criteria have been proposed to define "interesting," some of which are suitable for the classification problem. Each of these methods has undergone a variety of generalizations and tweaks to address a wider range of problems over the years.

The first *general* method to solve the classification problem automatically was the k-nearest neighbor approach [12]. k-Nearest neighbor estimates are known to be universally consistent if $k = k(n) \uparrow \infty$ slowly enough [21]. Their convergence, however, especially in high dimensional problems, can be slow in practice [36].

Local regression methods are a clever extension of this approach. To predict at a given point, instead of averaging the values given at the neighbors, they fit a low-order linear model to a locally-weighted version of the data set [10].

Trees [4] and neural nets [56] differ in that they search through a globally-parametrized class of functions. In all of these methods, cross-validation is often employed to estimate frequentist "out-of-sample" performance and select a regularization parameter which governs the trade-off between bias and variance [43].

Wavelet methods are in some ways a compromise between the local and the global approaches mentioned above. They fit an explicit global linear model to the data, but the basis elements in this model are carefully constructed to maintain "localization" (in space and frequency domains). They boast powerful asymptotic compression and approximation properties, computationally efficient transforms, and can employ special thresholding methods which "optimally" choose which coefficients in the model are kept [26]. However, their practical use seems to remain concentrated on the case of regularly-spaced regression data. Some recent papers address this shortcoming [14].

Support vector machines (SVMs) [65] employ a "kernel-trick" to reduce consideration of a certain globally-parametrized model class to consideration of an equivalent linear model class in an abstract Hilbert space. The estimated decision rule corresponds to the solution of a convex optimization problem. This objective function still involves an unknown regularization parameter. In practice, this parameter is often chosen by cross-validation, but, in principle, it can be chosen through consideration of the structural risk minimization (SRM) paradigm. The advantage of using the SRM paradigm is that one obtains provably valid confidence statements about the error rate that will obtain on future data. Moreover, these confidence bounds improve at an exponential rate in the number of data points. With realistic sample sizes, however, the bounds are often too crude to be of practical use. There are hidden connections between SVMs and (1) Bayesian methods employing Gaussian-process priors on the regression function (including the generalized spline methods of Wahba [66, 67]) (2) projection pursuit regression [11].

Bagging [3] and boosting [34] are meta-algorithms that "boost" the performance other classification algorithms (especially trees) by taking carefully chosen weighted averages of the results of the boosted (respectively, bagged) algorithm. There are close connections between boosting and the Lasso penalty, which itself is closely related to the least angle regression method (LARS) [30].

Chapter 3

Computing the Posterior

This chapter describes a Markov chain Monte Carlo (MCMC) algorithm that can be used to (approximately) draw samples from the posterior of the one-dimensional random step function prior π. This is essential, because for a complex prior like π, analytical evaluation of properties of the prior is intractable. All computation about the posterior, therefore, is made through (approximately) generating a large sample from it. Before describing these algorithms, it is natural to review the prior and then derive a more refined mathematical expression for the posterior. This exercise has the side effect of suggesting a more efficient sampling scheme. An informal sketch of the MCMC algorithm is then given in section 3.4. Additionally, section 3.5 explains an efficient way to use these samples to calculate the posterior mean. The interested reader is invited to download an implementation of these algorithms and others from the author's web page.

To define the algorithm more mathematically, a brief review of Markov chains and the Metropolis-Hastings algorithm for general state spaces is given. Section 3.7 gives a simple example of such a chain. Section 3.8 gives a mathematical treatment of the more complicated MCMC algorithm that was only sketched previously. It also verifies that the Markov chain satisfies detailed balance with respect to the posterior. This is done mainly to provide a thorough

and relatively self-contained theoretical analysis of the Markov chain.

Many of these calculations are essentially of the same type as those considered by Green [41]. Indeed, the birth-death move that I employ is essentially equivalent to the ones that he calls a "reversible jump." The verification of detailed balance is simple using his results. One need only observe that the transformations involved in these jumps essentially only permute coordinates; consequently the absolute value of the determinant of the Jacobian of this transformation is identically 1. Section 3.9 reviews two theoretical results that give sufficient conditions for a Markov chain to produce the intended ergodic sequence. Section 3.10 shows that these results are applicable so that, for example, the posterior mean that is computed will indeed approximate the intended curve.

3.1 Specification of the Prior in One Dimension

A review of notation and the prior is in order. Section 1.2 gives an informal description of the prior π and section 5.2 gives a more formal specification of the prior π and its parametrized version π'. The parameter space $\Theta = \cup_{k=1}^{\infty} \Theta_k$ where Θ_k parameterizes the class of functions $f : [0,1] \to [0,1]$ that have k locally-constant regions. Any such function is (essentially) determined by two vectors $\mathbf{s}$ and $\mathbf{v}$. Vector $\mathbf{s}$ lists the k values that the function takes on each region (as enumerated from left to right). Vector $\mathbf{v}$ lists (in no particular order) the $k-1$ locations at which the function jumps. This explains the definition:

$$\Theta_k := \{(k, \mathbf{v}, \mathbf{s}) \ : \ \mathbf{v} \in [0,1]^{k-1}, \mathbf{s} \in [0,1]^k\} \tag{3.1}$$

Θ_1 is a special case because there are no splits in a function that is everywhere constant. Define $\Theta_1 = \{(1, \langle\rangle, \mathbf{s}) \ : \ \mathbf{s} \in [0,1]^1\}$, where the *symbol* $\langle\rangle$ represents an empty list (which is not considered equivalent to an empty set). This is

consistent with the former definition of Θ_k if one allows the notation: $[0,1]^0 \equiv \{ \emptyset \}$.

To specify the prior π, first select a probability distribution κ on $\mathbb{N}$. κ represents the *a priori* distribution of the number of regions in the unknown function. In the introduction, the choice $\kappa \stackrel{d}{=} \mathit{Geometric}(\frac{1}{2})$ was suggested. Assume that $\kappa(k) > 0$ for all $k \in \mathbb{N}$, where (technically) $\kappa(k)$ is shorthand for $\kappa(\{k\})$. To pick a value $\theta \in \Theta$ from the (parametrized) prior π', first draw $K \sim \kappa$. Then, if $K = k$, draw $\mathbf{S} = \mathbf{s}$ uniformly from $[0,1]^k$ and draw $\mathbf{V} = \mathbf{v}$ uniformly from $[0,1]^{k-1}$. Form $\theta = (k, \mathbf{s}, \mathbf{v}) \in \Theta_k$. This completes the description of π'.

For a given point $\theta \in \Theta_k$ it is convenient associate a number of objects. Let $k(\theta)$, $\mathbf{v}(\theta)$, and $\mathbf{s}(\theta)$ stand for the k, $\mathbf{v}$, and $\mathbf{s}$ parts of θ respectively. Let $\mathbf{v}_{(i)}$ denote the i'th ordered value of $\mathbf{v}$. Additionally, for $\theta = (k, \mathbf{v}, \mathbf{s}) \in \Theta_k$ associate the function f_θ whose splits points and values are determined by $\mathbf{v}$ and $\mathbf{s}$. Specifically, for $k > 1$, let $I_1 = [0, \mathbf{v}_{(1)}), I_2 = [\mathbf{v}_{(2)}, \mathbf{v}_{(3)}), \ldots, I_{k-1} = [\mathbf{v}_{(k-2)}, \mathbf{v}_{(k-1)}), I_k = [\mathbf{v}_{(k-1)}, 1]$. If $k = 1$, just let $I_1 = [0,1]$. Take $f_\theta(x) = \sum_{i=1}^k \mathbf{s}_i \mathbf{1}_{x \in I_i}$.

I have chosen to work with uniform distributions on the splits and function values. Both of these choices could be varied, e.g. by using one-dimensional *Dirichlet* distributions for the split locations and using *Beta* distributions for the function values where, perhaps, the parameters of the *Beta* distribution are allowed to depend on spatial position. Diaconis and Freedman [25] adopt this level of generality (for the success probability prior), but I have not found it useful. What *would* be useful (but is avoided for simplicity of presentation) is to extend from binary classification to the multi-class case. This can be done by generalizing the $\mathit{Beta}(1,1)$ prior into a discrete *Dirichlet* prior on the class probabilities. Allowing dependencies among the parameters would more substantially complicate the analysis.

3.2 Representing the Posterior

The posterior is just the result of conditioning the prior on the data. The data is the list $\mathcal{D} := (x_1, y_1), \ldots, (x_n, y_n)$ where for i from 1 to n, $x_i \in [0, 1]$ represents where the i'th point occurred and $y_i \in \{0, 1\}$ represents whether the $Bernoulli(f(x_i))$ "coin" came up heads or tails. Denote the posterior distribution on θ given the data by $\pi'_\circ$:

$$\pi'_\circ(d\theta) := \pi'(d\theta \mid \mathcal{D}) \tag{3.2}$$

Provided that the denominator is non-zero and finite (which it will be) the posterior $\pi'_\circ$ has a density $\phi(\theta) = \phi(\theta; \mathcal{D})$ with respect to the prior π', so that $\pi'_\circ(d\theta) = \phi(\theta)\pi'(d\theta)$, where:

$$\phi(\theta) = \frac{L(\theta)}{\int_\Theta L(\theta')\pi'(d\theta')} \tag{3.3}$$

The likelihood function $L(\theta) = L(\theta; \mathcal{D})$ is defined by:

$$L(\theta) := \prod_{i=1}^{n} f_\theta(x_i)^{y_i} \left(1 - f_\theta(x_i)\right)^{1-y_i} \tag{3.4}$$

Recall that for $\theta = (k, \mathbf{v}, \mathbf{s})$, $f_\theta(x) = \sum_{j=1}^{k} \mathbf{s}_j \mathbf{1}_{x \in I_j}$, where the intervals I_j implicitly depend θ, but only through $\mathbf{v}$. To evaluate $f_\theta(x_i)$, then, is simply a matter of determining for which value j from 1 to k, $x_i \in I_j$ and then retrieving that $\mathbf{s}_j$. Call the value of j for which $x \in I_j$, $J(x; \theta)$. Then $f_\theta(x_i) = \mathbf{s}(\theta)_{J(x_i;\theta)}$. Consequently, L is simply a certain product of terms of the form $\mathbf{s}(\theta)_j$ or $1 - \mathbf{s}(\theta)_j$. To collect these together, define:

$$N_j^1 = N_j^1(\theta; \mathcal{D}) = (\#\text{ of data points in } I_j \text{ labeled 1}) \tag{3.5}$$

$$N_j^0 = N_j^0(\theta; \mathcal{D}) = (\#\text{ of data points in } I_j \text{ labeled 0}) \tag{3.6}$$

So that for $\theta \in \Theta_k$, (suppressing the dependence on θ and $\mathcal{D}$ from the right hand side) $L(\theta)$ can be expanded as:

$$L(\theta) = \prod_{j=1}^{k} \mathbf{s}_j^{N_j^1} (1 - \mathbf{s}_j)^{N_j^0} \tag{3.7}$$

Notice, then, that for fixed model number k and change-point locations $\mathbf{v}$, $L(\theta)$ only depends on the data through the (conditionally) sufficient statistics N_j^1 and N_j^0 for $j = 1, \ldots, k$. Moreover, for $\theta \in \Theta_k$, L is simply the product of k different binomial likelihood functions. This is intuitively obvious: If I have already decided exactly where the change-points are, the only remaining parameters are the success probabilities $\mathbf{s}$. Furthermore, according to the model, if I get a data-point (x, y), the x necessarily lands in some interval I_j and, then, the y is just the result of flipping an (independent) $\mathbf{s}_j$ coin.

Also notice that under the prior the $\mathbf{S}_j$'s are independent $U(0, 1)$ random variables. From the above discussion, it is apparent that if we condition on $K = k$ and $\mathbf{V} = \mathbf{v}$, the data simply tell us how many times each of the k "coins" with success probabilities $\mathbf{s}_1$ through $\mathbf{s}_k$ came up "heads" and "tails" as they were (collectively) flipped n times. Consequently, under the posterior, conditioned on $K = k$, $\mathbf{V} = \mathbf{v}$, and on the values of N_j^1 and N_j^0, each $\mathbf{S}_j$ is an independent $Beta(N_j^1 + 1, N_j^0 + 1)$ random variable. We recover this fact by direct calculation shortly.

These observations can be used to motivate an MCMC scheme that is substantially more efficient than the naive one that randomly changes k, $\mathbf{v}$, and $\mathbf{s}$ in the standard Metropolis-Hastings fashion. Namely, we will only have to use Markov Chain Monte Carlo steps in order to sample $(k, \mathbf{v})$ pairs from their marginal under the posterior. If desired, we can then create a complete sample, including a realization of $\mathbf{s}$ by sampling from the k independent *Beta* random variables whose parameters were explained above. To compute the posterior mean, $\mathbf{s}$ need not be simulated at all. The mean of the *Beta* distribution can be

computed analytically. This avoids substantial Monte Carlo error. For details, see sections 3.4 and 3.5.

Continuing with the calculations, write $\theta = (k, \mathbf{v}, \mathbf{s})$, and let $C = \{k\} \times A \times B = \{\theta \in \Theta_k : \mathbf{v} \in A, \mathbf{s} \in B\}$ for A, B measurable subsets of $[0,1]^{k-1}, [0,1]^k$ respectively. Compute:

$$\pi'_\circ(C) = \pi'(C|\mathcal{D}) = \int_C \phi(\theta)\pi'(d\theta) \tag{3.8}$$

$$= \frac{\int_C L(\theta)\pi'(d\theta)}{\int_\Theta L(\theta)\pi'(d\theta)} \tag{3.9}$$

$$\int_C L(\theta)\pi'(d\theta) = \kappa(k) \int_{\mathbf{v}\in A} \int_{\mathbf{s}\in B} L\left((k, \mathbf{v}, \mathbf{s})\right) d\mathbf{s}d\mathbf{v} \tag{3.10}$$

$$= \kappa(k) \int_{\mathbf{v}\in A} \int_{\mathbf{s}\in B} \prod_{j=1}^{k} \mathbf{s}_j^{N_j^1(\mathbf{v})} (1-\mathbf{s}_j)^{N_j^0(\mathbf{v})} d\mathbf{s}d\mathbf{v} \tag{3.11}$$

If, in particular, B is the rectangle $[a_1, b_1] \times \cdots \times [a_k, b_k]$ we get:

$$\int_C L(\theta)\pi'(d\theta) = \kappa(k) \int_{\mathbf{v}\in A} \left[\prod_{j=1}^{k} \int_{a_j}^{b_j} u^{N_j^1(\mathbf{v})} (1-u)^{N_j^0(\mathbf{v})} du \right] d\mathbf{v} \tag{3.12}$$

The inner integral is a *Beta* integral. Consider the special case in which $B = [0,1]^k$ (i.e. $C = \{k\} \times A \times [0,1]^k$). Then,

$$\int_C L(\theta)\pi'(d\theta) = \kappa(k) \int_{\mathbf{v}\in A} \rho_k(\mathbf{v}) d\mathbf{v} \tag{3.13}$$

Where $\rho_k(\mathbf{v})$ and also ρ_k are defined for $k \geq 2$ by:

$$\rho_k(\mathbf{v}) := \prod_{j=1}^{k} \frac{N_j^1(\mathbf{v})!\ N_j^0(\mathbf{v})!}{(N_j^1(\mathbf{v}) + N_j^0(\mathbf{v}) + 1)!} \tag{3.14}$$

$$\rho_k := \int_{\mathbf{v}\in[0,1]^{k-1}} \rho_k(\mathbf{v}) d\mathbf{v} \tag{3.15}$$

For $k = 1$, $\mathbf{v} = \langle\rangle$ and $I_1 = [0, 1]$ so that N_1^1 and N_1^0 are the total number of heads and tails respectively. In this case define $\rho_1 = \rho_1(\langle\rangle) = N_1^1!N_1^0!/(N_1^1 + N_1^0 + 1)!$.

The posterior probability of the k'th model is readily computed:

$$\pi'_\circ(\Theta_k) = \kappa(k)\rho_k/c \tag{3.16}$$

Where the normalizing constant c is the sum: $\sum_{j=1}^{\infty} \kappa(j)\rho_j$.

Now is a good time to notice that for any k and $\mathbf{v} \in [0, 1]^{k-1}$, and any data set $\mathcal{D}$, $\rho_k(\mathbf{v})$ and ρ_k are both positive and ≤ 1. Consequently, the same holds for c.

Let λ_j denote Lebesgue measure on $[0, 1]^j$. In the special case that $j = 0$, let λ_0 denote counting measure on the set $\{\langle\rangle\}$. Then the posterior density of the change-points $\mathbf{v}$ with respect to λ_{k-1}, given that model k holds is $\rho_k(\mathbf{v})/\rho_k$. Informally:

$$\pi'_\circ(\mathbf{V} \in d\mathbf{v} \mid K = k) = \rho_k(\mathbf{v})/\rho_k \lambda_j(d\mathbf{v}) \tag{3.17}$$

Finally, the posterior probability that $\mathbf{S}$ is in rectangle $B = [a_1, b_1] \times \cdots \times [a_k, b_k]$, given that model k holds and that the change-points are given by $\mathbf{v}$ is

indeed the same as k independent $Beta$'s:

$$\pi'_\circ(\mathbf{S} \in B \mid K = k, \mathbf{V} = \mathbf{v}) = \frac{1}{\rho_k(\mathbf{v})} \prod_{j=1}^{k} \int_{a_j}^{b_j} u^{N_j^1(\mathbf{v})}(1-u)^{N_j^0(\mathbf{v})} du \tag{3.18}$$

$$= \prod_{j=1}^{k} \mathbf{P}\left(Beta(N_j^1(\mathbf{v})+1, N_j^0(\mathbf{v})+1) \in [a_j, b_j]\right) du \tag{3.19}$$

Consequently, if $\widehat{s}_j$ denotes the expected value of $\mathbf{S}_j$ under the posterior, given that model k holds and that the change-points are given by $\mathbf{v}$, then $\widehat{s}_j$ is just the mean of the $Beta(N_j^1(\mathbf{v})+1, N_j^0(\mathbf{v})+1)$ distribution:

$$\widehat{s}_j = \frac{N_j^1(\mathbf{v})+1}{N_j^1(\mathbf{v})+N_j^0(\mathbf{v})+2} \tag{3.20}$$

3.3 Setup

This section sets up some basic definitions and ideas that underlie the algorithm described in the next section. The first definition, gives a new meaning to the symbol $\mathcal{X}$ which will be used throughout the remainder of this chapter. Elsewhere, $\mathcal{X}$ still stands for the covariate space. This should not introduce any confusion.

Write $\mathcal{X}_k$ for the parameter space formed from $(k, \mathbf{v})$ pairs with $\mathbf{v} \in [0,1]^{k-1}$ and build up the full parameter space $\mathcal{X}$ by taking the countable union:

$$\mathcal{X} := \cup_{k=1}^{\infty} \mathcal{X}_k \tag{3.21}$$

$$\mathcal{X}_k := \{(k, \mathbf{v}) \ : \ \mathbf{v} \in [0,1]^{k-1}\} \tag{3.22}$$

Again, $\mathcal{X}_1$ is a special case. Define $x_0 := (1, \emptyset)$ and let $\mathcal{X}_1 = \{x_0\}$, so that $\mathcal{X}_1$ is a singleton.

For convenience, let $k(x)$ stand for the k-part of x and let $\underline{x}$ stand for the $\mathbf{v}$-part of x. Bear in mind the nuisance that for $x \in \mathcal{X}_k$, $\underline{x} \in [0,1]^{k-1}$. Extend

this definition to sets; i.e. for a subset $C \subset \mathcal{X}_k$, define $\underline{C}$ as $\{\underline{x} \ : \ x \in C\}$.

For future reference, endow $\mathcal{X}$ with the σ-algebra $\mathcal{B}$ generated by sets of the form $k \times B$ where $k \geq 2$ and B is some (Borel) measurable subset of $[0,1]^{k-1}$ and by the singleton set $\mathcal{X}_1$. Let τ denote the extension of Lebesgue measure to $\mathcal{X}$. That is, for $k \geq 2$ and B a (Borel) measurable subset of $[0,1]^{k-1}$ define the τ measure of a set $C = k \times B \subset \mathcal{X}_k$, as the $k-1$-dimensional Lebesgue measure of $\underline{C} = B$. To account for $k = 1$, let $\tau(\mathcal{X}_1) = 1$.

Finally, combining the results in Equation 3.16 and Equation 3.17, the distribution under $\pi'_\circ$ of the random point $X = (K, V)$ in $\mathcal{X}$ can be computed. In the present notation, X has the density $\phi(x)$ with respect to τ:

$$\phi(x) := \kappa(k)\rho_k(\mathbf{v})/c \tag{3.23}$$

3.4 An MCMC Algorithm

This section contains a description of a (randomized) computational algorithm to produce a random sequence $\theta_1, \ldots, \theta_M$ drawn from the posterior distribution $\pi'_\circ$. A more formal version of this algorithm will be developed in section 3.8. Finally in section 3.10, it will be shown that the generated sequence has an ergodicity property. The main consequence of this is that if g is some measurable function with $\int |g(\theta)|\pi'_\circ(d\theta) < \infty$, one can use the average value of g on the sampled values to approximate the integral, in the sense that:

$$\frac{1}{M}\sum_{j=1}^{M} g(\theta_j) \to \int g(\theta)\pi'_\circ(d\theta) \tag{3.24}$$

The algorithm exploits the observation made in section 3.2 that under the posterior, conditioned on $K = k$, $\mathbf{V} = \mathbf{v}$, and on the values of N_j^1 and N_j^0, each $\mathbf{S}_j$ is an independent $Beta(N_j^1 + 1, N_j^0 + 1)$ random variable. Since $Beta$ random variables are easy to simulate and work with analytically, it suffices to simulate $(k, \mathbf{v})$ pairs from the (marginalized) posterior, instead of the full

$\theta = (k, \mathbf{v}, \mathbf{s})$.

Essentially, the MCMC algorithms that are developed in this chapter are a minor variation of the usual birth-death MCMC approach that is often used to simulate point processes. This approach is described by Geyer and Møller [38]. They claim geometric ergodicity for their Markov chain, under suitable restrictions on the sampling density. One technical distinction from these approaches is that when simulating a point process, the fundamental object is a subset of points $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}\}$; the theoretical treatment given in this chapter considers instead mathematical objects of the form $\big(k, (\mathbf{v}_1, \ldots, \mathbf{v}_{k-1})\big)$.

As far as the computations are concerned, though, there is no distinction between these formally different objects. Both could be represented on the computer *operationally* as a simple list of numbers called x with each number in the list specifying the location of a particular change-point. The algorithm merely assumes that it can call a function $\phi(x)$ that evaluates to the positive real number defined by Equation 3.23.

In particular, one can use $k(x)$ to find out that there are $k-1$ elements in this list. Furthermore the computer has no problem removing elements from the list all the way down to the empty list, or (ideally) adding elements one-by-one indefinitely. To agree with the notation of the previous section, if $k(x) = k$, then for $1 \leq j \leq k-1$, write $\underline{x}_j$ for the j'th element of the list. Write x_0 for the empty list.

Let M be the number of Monte Carlo samples that are desired. The algorithm simulates the workings of a Markov chain and generates the realizations: $x_1, x_2, \ldots, x_M$. It is assumed that for j from 1 to 5, p_j is a positive number, and that these numbers sum to one. These represent the mixture probabilities with with 5 component Markov chains are combined. For my computations, through a mixture of intuition and trial-and-error, I chose p as in Table 3.1. These values are by no means optimal. (The irregular numbers quoted here result from standardizing simpler ones so that they add to 1.)

p_1	p_2	p_3	p_4	p_5
0.1429	0.1429	0.2381	0.0476	0.4762

Table 3.1: *MCMC Mixture Probabilities*

Main MCMC Algorithm

Set $x = x_0$.
For i ranging from 1 to M, repeat the following steps:

1. Pick J at random from 1 to 5 with the probabilities p_1 through p_5 respectively.
2. Pick a "proposal" point y by following the subroutine specified in **Action J** (defined below).
3. Calculate $\alpha = \min(1, \phi(y)/\phi(x))$.
4. With probability α, set $x = y$.
5. Set $x_i = x$.

Action 1: Add or Delete a Random Coordinate

Set $y = x$.
Flip a fair coin.
If heads:
  Generate U uniformly at random on $[0, 1]$.
  Add U to the end of y.
  Return y.
If tails:
  Set $y = x$.
  If y is empty:
    Return y.
  Otherwise:
    Permute y randomly.
    Delete the last entry from y.
    Return y.
  End if
End if

Action 2: Randomize a Coordinate

Set $y = x$.
If y is empty:
  Return y.
Otherwise:
  Permute y randomly.
  Generate U uniformly at random on $[0, 1]$.
  Replace the last coordinate of y with U.
  Return y.
End if

Action 3: Shift a Coordinate

Set $y = x$.
If y is empty:
  Return y.
Otherwise:
  Permute y randomly.
  Set $U1 =$ the last element of y.
  Set $U2 = U1$+a random normal with mean 0 and standard deviation 0.1.
  If $U2$ is in $[0, 1]$:
    Replace the last coordinate of y with $U2$.
  End if
  Return y.
End if

Action 4: Randomize All Coordinates

Set $y = x$.
If y is empty:
  Return y.
Otherwise:
  For each coordinate of y do:
    Generate U uniformly at random on $[0, 1]$.
    Replace the current coordinate of y with U.
  End for
  Return y.
End if

Action 5: Shift All of the Coordinates Very Slightly

```
Set y = x.
If y is empty:
  Return y.
Otherwise:
  For each coordinate of y do:
    Set U1 = the current coordinate of y.
    Set U2 = U1+a random normal with mean 0 and standard deviation 0.01.
    If U2 is in [0, 1]:
      Replace the current coordinate of y with U2.
    Otherwise:
      Continue to the next coordinate.
    End if
  End for
  Return y.
End if
```

If desired, the resulting $x_1, \ldots, x_M$ can be randomly augmented into a sequence $\theta_1, \ldots, \theta_M$. To do so, simply generate all the necessary *Beta* random variables in order to sample $\mathbf{S}$ from its conditional distribution (c.f. Equation 3.18).

3.5 Posterior Mean Calculation

There are many potential uses of for the sample of θ values that can be approximately drawn from the posterior using the algorithm in the previous section. For example, for each sampled θ, a plot of the corresponding f_θ can be made, and inspecting some of these can give some idea about how confident to be about the shape of the unknown regression function.

This section, though, concentrates on estimating the mean of these functions. Call the resulting function the posterior mean, $\widehat{f}$. It represents the posterior's best estimate (in an $\mathcal{L}^2$ sense) of the unknown regression function. More formally, define the value of $\widetilde{f}$ at a point $u \in [0, 1]$ as the expected value

of $f_\theta(u)$ under the posterior on θ:

$$\widehat{f}(u) = \int_\Theta f_\theta(x)\pi'_\circ(d\theta) \tag{3.25}$$

$$= \int_{\mathcal{X}} \widehat{s}_{J(u;x)}\phi(x)\tau(dx) \tag{3.26}$$

Where $\widehat{s}_j$ was defined by Equation 3.20, $\phi(x)$ was defined by Equation 3.23, and τ was defined shortly before ϕ.

The algorithm from section 3.4 can be used to approximately generate a sample $x_1, \ldots, x_M$ from $\phi(x)\tau(dx)$ and then estimate this integral by:

$$\frac{1}{M}\sum_{i=1}^{M} \widehat{s}_{J(u;x)} \tag{3.27}$$

3.6 Metropolis-Hastings Markov Chains on General Spaces

Where does the algorithm described in section 3.4 come from? This section addresses the MCMC approach and begins a description of the larger framework within which algorithms like this one can be derived and evaluated.

Generally, MCMC techniques suggest how to formulate algorithms (more specifically Markov chains) that may be useful in order to sample a stationary ergodic sequence that converges to a given stationary distribution. This subject is very broad and active. For a review of the main ideas, see Tierney [64] or Liu [50]. MCMC techniques have opened up to numerical investigation a wide variety of Bayesian procedures, especially with the advent of the Gibbs sampler, the Metropolis-Hastings algorithm [51, 44], and its extension to the problem of "model determination" through "reversible jump" MCMC (Green [41]).

In very general terms (following [1]), the Markov chain setup is as follows. Let π be a probability distribution on a measurable space $(\mathcal{X}, \mathcal{B})$. Let P be a

transition probability function on this space, that is, P is a function on $\mathcal{X} \times \mathcal{B}$ such that, for each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$ and, for each $C \in \mathcal{B}$, $P(\cdot, C)$ is a measurable function on $(\mathcal{X}, \mathcal{B})$. The Markov chain $X_0, X_1, X_2, \ldots$ is generated as follows. We fix a starting point $X_0 = x_0$, generate an observation X_1 from $P(X_0, \cdot)$, generate an observation X_2 from $P(X_1, \cdot)$, and so on.

P will be constructed to obey detailed balance with respect to π. Namely:

$$\int_A P(x, B)\pi(dx) = \int_B P(x, A)\pi(dx) \text{ for every } A, B \in \mathcal{B} \tag{3.28}$$

This condition is very convenient because although it will be easy to construct chains that satisfy it, it is also powerful. In particular, (by choosing $B = \mathcal{X}$) it implies that π is *an* invariant measure for the Markov chain, that is,

$$\pi(A) = \int_{\mathcal{X}} P(x, A)\pi(dx) \text{ for every } A \in \mathcal{B} \tag{3.29}$$

The goal is to choose P so that π is the *unique* invariant measure; and, moreover that the Markov chain will produce an ergodic sequence of observations from π. For this goal, detailed balance is a useful (although not necessary) "first step."

Suppose $\widetilde{P}$ is some transition probability function which (presumably) does not satisfy reversibility with respect to π. Let $\widetilde{\mu}(dx, dy) := \pi(dx)\widetilde{P}(x, dy)$. Suppose that $\widetilde{\mu}$ is absolutely continuous with respect to some symmetric σ-finite measure μ. Specifically, suppose that μ is a measure on the measurable space $(\mathcal{X} \times \mathcal{X}, \sigma(\mathcal{B} \times \mathcal{B}))$ that satisfies $\mu(A \times B) = \mu(B \times A)$ for all $A, B \in \mathcal{B}$. In simple cases, one can choose μ to be a product measure; for example $\pi(dx)\pi(dy)$ or $\lambda(dx)\lambda(dy)$ where perhaps π has a density with respect to λ. If no such measure is readily available, one may take $\mu(dx, dy) = \frac{1}{2}\widetilde{\mu}(dx, dy) + \frac{1}{2}\widetilde{\mu}(dy, dx)$; i.e. $\mu(A \times B) = \frac{1}{2}\widetilde{\mu}(A \times B) + \frac{1}{2}\widetilde{\mu}(B \times A)$ for $A, B \in \mathcal{B}$. Let $\widetilde{p}$ be a version of the Radon-Nikodým derivative of $\widetilde{\mu}$ with respect to μ so that $\widetilde{\mu}(dx, dy) =$

$\widetilde{p}(x,y)\mu(dx,dy)$. If $\widetilde{p}(x,y)$ happens to be symmetric; i.e. $\widetilde{p}(x,y) = \widetilde{p}(y,x)$ for all $x, y \in \mathcal{X}$, then $\widetilde{P}$ already satisfies detailed balance with respect to π. To verify this, compute that for every every $A, B \in \mathcal{B}$:

$$\begin{aligned}
\int_A \widetilde{P}(x,B)\pi(dx) &= \int_{x\in A}\int_{y\in B} \pi(dx)P(x,dy) && (3.30)\\
&= \int_{(x,y)\in A\times B} \widetilde{p}(x,y)\mu(dx,dy) && (3.31)\\
&= \int_{(x,y)\in A\times B} \widetilde{p}(y,x)\mu(dx,dy) && (3.32)\\
&= \int_{(x,y)\in B\times A} \widetilde{p}(x,y)\mu(dx,dy) && (3.33)\\
&= \int_{x\in B}\int_{y\in A} \pi(dx)P(x,dy) && (3.34)\\
&= \int_B \widetilde{P}(x,A)\pi(dx) && (3.35)
\end{aligned}$$

When $\widetilde{p}$ is not symmetric, there is no trouble in constructing the closely related symmetric function $p(x,y) = \min(\widetilde{p}(x,y), \widetilde{p}(y,x))$. Does this suggest how to construct a probability transition function P based on $\widetilde{P}$ that satisfies detailed balance with respect to π? Yes, fortunately it does. Define:

$$\alpha(x,y) = \begin{cases} \min(1, \widetilde{p}(y,x)/\widetilde{p}(x,y)) & \text{if } \widetilde{p}(x,y) > 0 \\ 1 & \text{if } \widetilde{p}(x,y) = 0 \end{cases} \tag{3.36}$$

Then (check) $p(x,y) = \alpha(x,y)\widetilde{p}(x,y)$ for all $x, y \in \mathcal{X}$. This suggests defining $Q(x,dy) = \alpha(x,y)\widetilde{P}(x,dy)$, so that $\pi(dx)Q(x,dy) = \alpha(x,y)\widetilde{p}(x,y)\mu(dx,dy) = p(x,y)\mu(dx,dy)$. The only problem is that $Q(x,\cdot)$ is not a probability (in general), but a sub-probability. To account for the "forgotten" mass, set $h(x) = 1 - Q(x,\mathcal{X})$ for all $x \in \mathcal{X}$. And define $P(x,dy) = Q(x,dy) + h(x)\delta_x(dy)$. Here,

$\delta_x(\cdot)$ stands for the Dirac measure at x. In expanded form, this defines:

$$P(x, dy) = \alpha(x, y)\widetilde{P}(x, dy) + \left[\int_{\mathcal{X}} (1 - \alpha(x, z))\widetilde{P}(x, dz)\right] \delta_x(dy) \qquad (3.37)$$

In conclusion, one can now easily verify that this defines a probability transition function P which satisfies detailed balance with respect to π and which is a simple modification of $\widetilde{P}$. Indeed, P is simply a modification of $\widetilde{P}$ that sometimes "holds" instead of taking the transition that $\widetilde{P}$ proposes. Suppose that $Y = y$ is a particular value drawn from $\widetilde{P}(x, \cdot)$. That is, suppose that $\widetilde{P}$ has "proposed" the transition from x to y. Then P "accepts" this transition with probability $\alpha(x, y)$, but holds with probability $1 - \alpha(x, y)$. Furthermore, because $\alpha(x, y)$ only depends on the ratio $\widetilde{p}(y, x)/\widetilde{p}(x, y)$ it is sufficient to be able to compute $\widetilde{p}(x, y)$ up to an unknown constant factor.

Clearly, then, P is a computationally simple modification of $\widetilde{P}$; the only caveat is that $\alpha(x, y)$ may often be very small or even 0 so that in the extreme degenerate case in which $\alpha \equiv 0$, P is the Markov chain that always holds. Indeed, this Markov chain is reversible with respect to any distribution, but it certainly does not serve the larger goal of producing an ergodic sequence of realizations from π. Similar problems can occur if $\widetilde{P}$ is not transitive or is otherwise unsuitable. For these reasons the ergodicity conditions from section 3.9 are needed.

3.7 A Simple Markov Chain

A simple example is in order to make these ideas more concrete. This chain will not be as efficient (in practice) as the local-move Markov chain developed in the next section.

In words, this will be the chain that stays fixed (holds) at its current value $x \in \mathcal{X}$ until a new value $y\mathcal{X}$ drawn from the prior is accepted; y will always be accepted if y makes the data more likely (i.e. higher predictive probability

under our model); otherwise it is accepted with a probability reflecting the ratio of the predictive probabilities. In this way, the chain readily walks "uphill," but, with just the right probability (because of the detailed balance condition that will be shown) it also walks "downhill."

Recall the notation from section 3.3 that defined the measurable space $(\mathcal{X}, \mathcal{B})$ and denote the posterior distribution on this space by $\pi'_\circ$. For convenience, recall that for any $x \in \mathcal{X}$, $x = (k, \mathbf{v})$ and write $\kappa(x) = \kappa(k(x))$ and $\rho(x) = \rho(\mathbf{v}(x))$ so that the prior π on points $y \in \mathcal{X}$ has density $\kappa(y)$ with respect to τ. For any $x \in \mathcal{X}$ and any $B \in \mathcal{B}$, let $\widetilde{P}$ be defined by $\widetilde{P}(x, B) = \int_{y \in B} \kappa(y)\tau(dy)$. That is, $\widetilde{P}$ is the probability transition function that (without reference to x) samples a new point $y \in \mathcal{X}$ from the prior. Now expand $\pi'_\circ(dx)P(x, dy) = \phi(x)\kappa(y)\tau(dx)\tau(dy) = (\rho(x)/c)\kappa(x)\kappa(y)\tau(dx)\tau(dy)$. Conveniently then, this distribution has a density with respect to the product measure $\kappa(x)\tau(dx)\kappa(y)\tau(dy)$. To agree with the notation in the previous section, let μ denote this product measure and let $\widetilde{p}(x, y) = \rho(x)/c$. Recall that each of these terms is always positive. Because of this and a convenient cancellation, the expression for α becomes

$$\alpha(x, y) = \min(1, \rho(y)/\rho(x))$$

Finally, as before, define $Q(x, dy) = \alpha(x, y)\widetilde{P}(x, dy)$, $h(x) = 1 - Q(x, \mathcal{X})$, and set $P(x, dy) = Q(x, dy) + h(x)\delta_x(dy)$.

3.8 A Local-Move Markov Chain

This section gives a formal definition of the Markov chain type algorithm that was explained in section 3.4. It then shows that this chain satisfies detailed balance with respect to $\pi'_\circ$. To introduce this more useful, but more complicated Markov chain on $(\mathcal{X}, \mathcal{B})$, some notation is needed. When $j \geq 2$, and $\mathbf{v} \in [0, 1]^j$, write $\mathbf{v}_-$ for the vector $(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}) \in [0, 1]^{j-1}$ which leaves off the

last coordinate of $\mathbf{v}$. Consider some point $x = (k, \mathbf{v}) \in \mathcal{X}$. Let $g_\downarrow(x)$ stand for the point in $\mathcal{X}$ that is "one level down" from x with the last coordinate of $\mathbf{v}$ having been removed. That is, when $k \geq 3$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}) \in [0,1]^{k-1}$, let $g_\downarrow(x) = (k-1, \mathbf{v}_-)$. For $x = (2, (\mathbf{v}_1)) \in \mathcal{X}_2$, let $g_\downarrow(x) = (1, \emptyset)$. For $x = (1, \emptyset) \in \mathcal{X}_1$, there is no further down to go, and so let $g_\downarrow(x) = x$. Let $\lambda^x_\downarrow(\cdot) = \delta_{g_\downarrow(x)}(\cdot)$ denote the Dirac measure at $g_\downarrow(x)$.

Similarly, let $g_\uparrow(x, u)$ stand for the point in $\mathcal{X}$ that is "one level up" from x, where the last coordinate is filled in with u. That is, for $k \geq 2$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_{k-1})$, $g_\uparrow(x, u) = (k+1, (\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}, u))$. For $x = (1, \emptyset)$, let $g_\uparrow(x, u) = (2, (u))$. Let $\lambda^x_\uparrow(\cdot)$ denote the distribution of $g_\uparrow(x, U)$ where U is uniformly distributed on $[0, 1]$.

Let $\lambda_j(\cdot)$ denote Lebesgue measure on $\mathbb{R}^j$ and for $B \in \mathcal{B}$ where $B \subset \mathcal{X}_k$, let $\underline{B} = \{\mathbf{v} \in [0,1]^{k-1} \ : \ (k, \mathbf{v}) \in B\}$.

To define the transition probability function P on $(\mathcal{X}, \mathcal{B})$ satisfying detailed balance with respect to $\pi'_\circ$, I first define various transition probability functions $\widetilde{P}_j(x, dy)$; then, for a generic function $\alpha_j(x, y)$, define:

$$Q_j(x, dy) = \alpha_j(x, dy)\widetilde{P}_j(x, dy) \tag{3.38}$$

Next the α_j's are chosen so that for every j, Q_j satisfies the appropriate detailed balance formula:

$$\int_{x \in A} \pi(dx) \int_{y \in B} Q_j(x, dy) = \int_{x \in B} \qquad \pi(dx) \int_{y \in A} Q_j(x, dy) \tag{3.39}$$

It is easily verified then that $P_j = Q_j(x, dy) + [1 - Q_j(x, \mathcal{X})]\delta_x(dy)$, is a transition probability which satisfies detailed balance. Finally, set:$P(x, dy) = \sum_j p_j P_j(x, dy)$.

Generally the proposals I consider are "symmetric" and so $\alpha_j(x, y)$ will work out to be $\min(1, \phi(y)/\phi(x))$ in each case. An exception is in Equation 3.86, but when the proposal density is simple, (as it is in the case of interest) it also

reduces to the previous case.

To begin, set $\widetilde{P}_1(x, dy) = \frac{1}{2}\lambda^x_\uparrow(dy) + \frac{1}{2}\lambda^x_\downarrow(dy)$. This transition probability function represents the chain that adds or deletes coordinates from the vector $\mathbf{v}$ randomly.

To choose α_1, first compute the left and right hand sides of the detailed balance equation for Q_1. Let $A, B \in \mathcal{B}$ with $A \subset \mathcal{X}_j$ and $B \subset \mathcal{X}_k$. There are three cases of interest for j and k: (1) $k = j+1, j \geq 2$, (2) $j = 1, k = 2$, (3) $j = 1, k = 1$. These account for all the possibilities because if $j < k$, simply replace the roles of j and k; if $k > j+1$ or $j = k \neq 1$ both sides will evaluate to 0. If α_1 can be chosen to set the left and right sides equal for every such case, detailed balance is proven because general $A, B \in \mathcal{B}$ can be decomposed into these component subsets.

Suppose that $k = j + 1$, and calculate:

$$\int_{x\in A} \pi(dx) \int_{y\in B} Q_1(x, dy) \tag{3.40}$$

$$= \int_{x\in A}\int_{y\in B} \phi(x)\alpha_1(x, y)\tau(dx)[\frac{1}{2}\lambda^x_\uparrow(dy) + \frac{1}{2}\lambda^x_\downarrow(dy)] \tag{3.41}$$

$$= \int_{x\in A}\int_{y\in B} \frac{1}{2}\phi(x)\alpha_1(x, y)\tau(dx)\lambda^x_\uparrow(dy) \tag{3.42}$$

$$= \int_{\mathbf{v}\in \underline{A}}\int_{u\in C(\mathbf{v};B)} \frac{1}{2}\kappa(j)\rho_j(\mathbf{v})\alpha_1((j, \mathbf{v}), g_\uparrow((j, \mathbf{v}), u))\lambda_{j-1}(d\mathbf{v})du \tag{3.43}$$

$$= \int_{\mathbf{w}\in \underline{D}} \frac{1}{2}\kappa(j)\rho_j(\mathbf{w}_-)\alpha_1((j, \mathbf{w}_-), g_\uparrow((j, \mathbf{w}_-), u))\lambda_j(d\mathbf{w}) \tag{3.44}$$

$$= \int_{x\in D} \frac{1}{2}\phi(g_\downarrow(x))\alpha_1(g_\downarrow(x), x)\tau(dx) \tag{3.45}$$

Where:

$$C(\mathbf{v}; B) = \{u \in [0, 1] \ : \ g_\uparrow((j, \mathbf{v}), u) \in B\} \tag{3.46}$$

$$\underline{D} = \{\mathbf{w} \in [0,1]^j \ : \ \mathbf{w}_- \in \underline{A}, w_j \in C(\mathbf{w}_-; B)\} \tag{3.47}$$

$$= \{\mathbf{w} \in \underline{B} \ : \ \mathbf{w}_- \in \underline{A}\} \tag{3.48}$$

$$D = \{x = (k, \mathbf{w}) \in C_k \ : \ \mathbf{w} \in \underline{D}\} \tag{3.49}$$

$$= \{x \in B \ : \ g_\downarrow(x) \in A\} \tag{3.50}$$

Similarly, compute:

$$\int_{x \in B} \pi(dx) \int_{y \in A} Q_1(x, dy) \tag{3.51}$$

$$= \int_{x \in B} \int_{y \in A} \phi(x) \alpha_1(x, y) \tau(dx) [\frac{1}{2} \lambda^x_\uparrow(dy) + \frac{1}{2} \lambda^x_\downarrow(dy)] \tag{3.52}$$

$$= \int_{x \in B} \int_{y \in A} \frac{1}{2} \phi(x) \alpha_1(x, y) \tau(dx) \lambda^x_\downarrow(dy) \tag{3.53}$$

$$= \int_{x \in B} \frac{1}{2} \phi(x) \alpha_1(x, g_\downarrow(x)) \tau(dx) \mathbf{1}_{g_\downarrow(x) \in A} \tag{3.54}$$

$$= \int_{x \in D} \frac{1}{2} \phi(x) \alpha_1(x, g_\downarrow(x)) \tau(dx) \tag{3.55}$$

Suppose that $j = 1, k = 2$, so that $A = \{x_0\}$ or $A =$ where $x_0 = (1, \langle\rangle)$ and

any $x \in B$ is of the form $(2, (\mathbf{v}_1))$. Then:

$$\int_{x\in A} \pi(dx) \int_{y\in B} Q_1(x, dy) \tag{3.56}$$

$$= \int_{x\in A} \int_{y\in B} \phi(x)\alpha_1(x, y)\tau(dx)[\frac{1}{2}\lambda^x_\uparrow(dy) + \frac{1}{2}\lambda^x_\downarrow(dy)] \tag{3.57}$$

$$= \int_{x\in A} \int_{y\in B} \frac{1}{2}\phi(x)\alpha_1(x, y)\tau(dx)\lambda^x_\uparrow(dy) \tag{3.58}$$

$$= \int_{y\in B} \frac{1}{2}\mathbf{1}x_0 \in A\phi(x_0)\alpha_1(x_0, y)\lambda^{x_0}_\uparrow(dy) \tag{3.59}$$

$$= \int_{\mathbf{v}_1\in \underline{B}} \frac{1}{2}\mathbf{1}x_0 \in A\phi(x_0)\alpha_1(x_0, (2, \mathbf{v}_1))d\mathbf{v}_1 \tag{3.60}$$

$$= \int_{x\in D} \frac{1}{2}\phi(x_0)\alpha_1(x_0, x)\tau(dx) \tag{3.61}$$

Similarly, compute:

$$\int_{x\in B} \pi(dx) \int_{y\in A} Q_1(x, dy) \tag{3.62}$$

$$= \int_{x\in B} \int_{y\in A} \phi(x)\alpha_1(x, y)\tau(dx)[\frac{1}{2}\lambda^x_\uparrow(dy) + \frac{1}{2}\lambda^x_\downarrow(dy)] \tag{3.63}$$

$$= \int_{x\in B} \int_{y\in A} \frac{1}{2}\phi(x)\alpha_1(x, y)\tau(dx)\lambda^x_\downarrow(dy) \tag{3.64}$$

$$= \int_{x\in B} \frac{1}{2}\phi(x)\alpha_1(x, g_\downarrow(x))\tau(dx)\mathbf{1}_{g_\downarrow(x)} \in A \tag{3.65}$$

$$= \int_{x\in D} \frac{1}{2}\phi(x)\alpha_1(x, x_0)\tau(dx) \tag{3.66}$$

The only remaining case to compute is where $j = k = 1$ and the only case of interest here is the one in which $A = B = \{x_0\}$. Even this is trivial, because the left and right sides of this symmetric case must surely match.

Recalling that $\phi(x) > 0$ for all $x \in \mathcal{X}$, for $x, y \in \mathcal{X}$ detailed balance is

achieved upon defining:

$$\alpha_1(x, y) = \min(\phi(y)/\phi(x), 1) \tag{3.67}$$

To define $\widetilde{P}_2$ additional notation is needed. For $k \geq 2, 1 \leq j \leq k-1$ let $\lambda^x_{\cdot j}(\cdot)$ denote the distribution on $(\mathcal{X}, \mathcal{B})$ under which the j'th coordinate of the $\mathbf{v}$ in $x = (k, \mathbf{v})$ is replaced with a uniform random variable. For $k \geq 2$ and $x \in \mathcal{X}_k$, let $\lambda^x_{\cdot}(dy) = \frac{1}{k-1}\sum_{j=1}^{k-1} \lambda^x_{\cdot j}(dy)$, i.e. the probability distribution that picks a coordinate of the $\mathbf{v}$ in x randomly and then changes it to a uniformly random value. For $x = x_0$, let $\lambda^x_{\cdot}(dy) = \delta_{x_0}(dy)$.

Now define $\widetilde{P}_2(x, dy) = \lambda^x_{\cdot}(dy)$. More generally, define:

$$\widetilde{P}_3(x, dy) = \Psi(x, y)\lambda^x_{\cdot}(dy) + h_\Psi(x)\delta_x(dy) \tag{3.68}$$

Where Ψ is any (measurable) non-negative function on $\mathcal{X} \times \mathcal{X}$ satisfying for all $x, y \in \mathcal{X}$: (1) $\Psi(x, y) = \Psi(y, x)$, (2) $\int_{\mathcal{X}_k} \Psi(x, y)\lambda^x_{\cdot}(dy) + h_\Psi(x) = 1$ for some non-negative $h_\Psi(x)$. For example, if $x = (k, \mathbf{v}_x)$, $y = (k, \mathbf{v}_y)$, and ϕ_σ represents the univariate normal density with standard deviation $\sigma > 0$, take $\Psi(x, y) = \phi_\sigma(||\mathbf{v}_x - \mathbf{v}_y||_2)$ and $h_\Psi(x) = 1 - \int_{\mathcal{X}_k} \Psi(x, y)\lambda^x_{\cdot}(dy)$. Fortunately, there will be no need to compute h_Ψ; it is enough to note that it is non-negative.

To check the detailed balance for Q_3 it is sufficient to consider $A, B \in \mathcal{B}$ where $A, B \subset \mathcal{X}_k$.

$$\int_{x\in A}\pi(dx)\int_{y\in B}Q_3(x,dy) \tag{3.69}$$

$$= \int_{x\in A}\int_{y\in B}\phi(x)\alpha_3(x,y)\tau(dx)[\Psi(x,y)\lambda_{\cdot}^{x}(dy)+h_\Psi(x)\delta_x(dy)] \tag{3.70}$$

$$\begin{aligned}= &\int_{x\in A}\int_{y\in B}\phi(x)\alpha_3(x,y)\Psi(x,y)\tau(dx)\lambda_{\cdot}^{x}(dy)\\ &\quad+\int_{x\in A}\phi(x)\alpha_3(x,x)h_\Psi(x)\mathbf{1}_{x\,\in\, B}\tau(dx)\end{aligned} \tag{3.71}$$

$$\begin{aligned}= &\int_{x\in A}\int_{y\in B}\phi(x)\alpha_3(x,y)\Psi(x,y)\tau(dx)\lambda_{\cdot}^{x}(dy)\\ &\quad+\int_{x\in A\cap B}\phi(x)\alpha_3(x,x)h_\Psi(x)\tau(dx)\end{aligned} \tag{3.72}$$

This second term in the latter expression does not change if we replace A, B with B, A and consequently needs no further consideration. To expand the first term, recall that in this situation $\lambda_{\cdot}^{x}$ just picks one of the $k-1$ coordinates and randomizes it. Let $S_j^k(\mathbf{w})$ stand for the modified version of $\mathbf{w}$ in which the j'th and k'th coordinates of $\mathbf{w}$ are swapped. In suggestive notation, let:

$$\widetilde{X} = \widetilde{X}(\mathbf{w}) = (k, w_-) \tag{3.73}$$

$$\widetilde{Y} = \widetilde{Y}(\mathbf{w}) = \widetilde{X}(S_j^k(\mathbf{w})) \tag{3.74}$$

$$D_j = \{\mathbf{w}\in[0,1]^k \;:\; \widetilde{X}(\mathbf{w})\in A,\ \widetilde{Y}(\mathbf{w})\in B\} \tag{3.75}$$

$$D_j' = \{\mathbf{w}\in[0,1]^k \;:\; \widetilde{X}(\mathbf{w})\in B,\ \widetilde{Y}(\mathbf{w})\in A\} \tag{3.76}$$

Notice that $\mathbf{w}\in D_j \iff S_j^k(\mathbf{w})\in D_j'$.

Then, suppressing the $\mathbf{w}$ dependence from $\widetilde{X}$ and $\widetilde{Y}$, the first term expands

into:

$$\frac{1}{k-1}\sum_{j=1}^{k-1}\int_{\mathbf{w}\in D_j}\phi(\widetilde{X})\alpha_3(\widetilde{X},\widetilde{Y})\Psi(\widetilde{X},\widetilde{Y})\lambda_k(d\mathbf{w}) \tag{3.77}$$

$$=\frac{1}{k-1}\sum_{j=1}^{k-1}\int_{\mathbf{w}\in D'_j}\phi(\widetilde{Y})\alpha_3(\widetilde{Y},\widetilde{X})\Psi(\widetilde{Y},\widetilde{X})\lambda_k(d\mathbf{w}) \tag{3.78}$$

$$=\frac{1}{k-1}\sum_{j=1}^{k-1}\int_{\mathbf{w}\in D'_j}\phi(\widetilde{Y})\alpha_3(\widetilde{Y},\widetilde{X})\Psi(\widetilde{X},\widetilde{Y})\lambda_k(d\mathbf{w}). \tag{3.79}$$

In summary then, (using Equation 3.79)

$$\int_{x\in A}\pi(dx)\int_{y\in B}Q_3(x,dy) \tag{3.80}$$

$$=\frac{1}{k-1}\sum_{j=1}^{k-1}\int_{\mathbf{w}\in D'_j}\phi(\widetilde{Y})\alpha_3(\widetilde{Y},\widetilde{X})\Psi(\widetilde{X},\widetilde{Y})\lambda_k(d\mathbf{w})+const \tag{3.81}$$

While, using Equation 3.77 with the roles of A and B interchanged so that "D_j"=D'_j:

$$\int_{x\in B}\pi(dx)\int_{y\in A}Q_3(x,dy) \tag{3.82}$$

$$=\frac{1}{k-1}\sum_{j=1}^{k-1}\int_{\mathbf{w}\in D'_j}\phi(\widetilde{X})\alpha_3(\widetilde{X},\widetilde{Y})\Psi(\widetilde{X},\widetilde{Y})\lambda_k(d\mathbf{w})+const \tag{3.83}$$

Again we achieve balance using:

$$\alpha_3(x,y)=\min(\phi(y)/\phi(x),1) \tag{3.84}$$

Define $\widetilde{P}_4(x,dy)=\xi_x(y)\tau(dy)$. Where for any $x\in\mathcal{X}$, $\xi_x(\cdot)$ is a density with respect to τ. Accordingly:

$$\pi'_\circ(dx)\widetilde{P}_4(x,dy)=\phi(x)\xi_x(y)\tau(dx)\tau(dy)=\widetilde{p}(x,y)\tau(dx)\tau(dy) \tag{3.85}$$

By the arguments in section 3.6 the corresponding transition probability function P_4 satisfies detailed balance with respect to $\pi'_\circ$ provided that:

$$\alpha_4(x,y) = \begin{cases} \min(1, [\phi(y)\xi_y(x)]/[\phi(x)\xi_x(y)] & \text{if } \phi(x)\xi_x(y) > 0 \\ 1 & \text{if } \phi(x)\xi_x(y) = 0 \end{cases} \tag{3.86}$$

For example, for $x \in \mathcal{X}_k$, choose $\xi_x(y) = \mathbf{1}_{y \in \mathcal{X}_k}$. And for this ξ_x, the familiar choice $\alpha_4(x,y) = \min(1, \phi(y)/\phi(x))$ works equally well.

Finally, define $\widetilde{P}_5(x, dy)$ for $x \in \mathcal{X}_k$ as the composition of the $k-1$ separate transition probability functions $\widetilde{P}_5^j$ applied sequentially from $j = 1$ to $j = k-1$. Each $\widetilde{P}_5^j$ is just intended to move the j'th coordinate of $\mathbf{v}_x$ by a small amount (or hold). In effect, then, $\widetilde{P}_5$ moves all of the coordinates of x randomly, but technically speaking some subset of them may hold on any given step. For $x \in \mathcal{X}_k$ and $1 \leq j \leq k-1$ set $\widetilde{P}_5^j(x, dy) = \Psi(x,y)\lambda_{\cdot j}^x(dy) + h_\Psi(x)\delta_x(dy)$ for some Ψ and h_Ψ satisfying the same constraints as made when defining $\widetilde{P}_3$. I omit the verification that by choosing $\alpha_5 = \min(1, \phi(y)/\phi(x))$, as usual, P_5 satisfies detailed balance.

This concludes a consideration of each chain P_1 through P_5. Each was shown to satisfy detailed balance with respect to $\pi'_\circ$. Consequently, their mixture P also satisfies detailed balance with respect to $\pi'_\circ$.

As an aside, notice that one may compose any of the P_j with a random permutation since ϕ itself is invariant with respect to permuted $\mathbf{v}$. Doing so will allow P_1 to add and (by pre-composing) delete coordinates in arbitrary locations.

3.9 Markov Chain Convergence Theory

Early results on the ergodicity of Markov chains on general state spaces used a condition known as the Doeblin condition. It implies that there exists an

invariant probability measure to which the Markov chain converges at a geometric rate, from any starting point.[1]

Theorem 1 (Doob (1953) [28]). *Suppose that the Markov chain on the measure space $(\mathcal{X}, \mathcal{B})$ generated by probability transition function P satisfies the Doeblin condition that there is a probability measure λ on $(\mathcal{X}, \mathcal{B})$, an integer k, and an $\epsilon > 0$ such that:*

$$P^k(x, C) \geq \epsilon\lambda(C) \quad \text{for all } x \in \mathcal{X} \text{ and all } C \in \mathcal{B}$$

Then there exists a unique invariant probability measure π such that for all n,

$$\sup_{C \in \mathcal{B}} |P^n(x, C) - \pi(C)| \leq (1 - \epsilon)^{(n/k)-1} \qquad \text{for all } x \in \mathcal{X}$$

An easy corollary is that if P satisfies the conditions of the theorem, and was already known to be reversible with respect to some specific distribution, then that same distribution must be the unique stationary distribution π. The Doeblin condition is quite strong and rarely holds in applications.

Athreya, Doss, and Sethuraman [1] prove an ergodicity result for general state spaces whose conditions hold much more broadly and remain reasonably easy to check. An abbreviated version is given below

Theorem 2 (Athreya, Doss, and Sethuraman (1996)). *Suppose that the Markov chain $\{X_n\}$ with transition function $P(x, C)$ has an invariant probability measure π, that is Equation 3.29 holds. Suppose that there is a set $A \in \mathcal{B}$, a probability measure λ with $\lambda(A) = 1$, a constant $\epsilon > 0$, and an integer $n_0 \geq 1$ such that:*

$$\pi(\{x : P_x(X_m \in A \text{ for some } m \geq 1) > 0\}) = 1 \tag{3.87}$$

[1] This section reviews some material given in a paper by Athreya, Doss, and Sethuraman [1]

and

$$P^{n_0}(x,\cdot) \geq \epsilon\lambda(\cdot) \quad \textit{for each } x \in A \tag{3.88}$$

Further suppose that either $n_0 = 1$ *or that*

$$\text{g.c.d.}\{m \ : \ \textit{there is an } \epsilon_m > 0 \textit{ such that } P^m(x,\cdot) \geq \epsilon_m\lambda(\cdot) \textit{ for each } x \in A\} = 1 \tag{3.89}$$

Then there is a set $D \in \mathcal{B}$ *such that*

$$\pi(D) = 1 \quad \textit{and} \quad \sup_{C\in\mathcal{B}} |P^n(x,C) - \pi(C)| \to 0 \ \ \textit{for each } x \in D. \tag{3.90}$$

Let $f(x)$ *be a measurable function on* $(\mathcal{X},\mathcal{B})$ *such that* $\int \pi(dy)|f(y)| < \infty$. *Then*

$$P_x\left\{\frac{1}{n}\sum_{j=1}^{n} f(X_j) \to \int \pi(dy)f(y)\right\} = 1 \quad \textit{for } [\pi]\textit{-almost all } x \tag{3.91}$$

3.10 Convergence Results

A short proof suffices to show that theorem 2 applies to the local-move Markov chain. It is assumed that mixing probability $p_1 > 0$. The proof takes advantage of the atom $x_0 = \big(1, \emptyset\big)$.

Let P denote the local-move Markov chain from section 3.8. It was already verified there that P satisfies detailed balance with respect to $\pi'_\circ$, and it follows that P has invariant probability measure $\pi'_\circ$. Let $A = \mathcal{X}_1$, the singleton set $\{x_0\}$. Let $\lambda(\cdot)$ be counting measure on $\mathcal{X}_1$. Let $n_0 = 1$. Set $\epsilon = P(x_0, \mathcal{X}_1)$, i.e. the chance of holding at this atom. This quantity is positive (because $P(x_0,\mathcal{X}_1) \geq p_1 P_1(x_0,\{x_0\}) \geq p_1\widetilde{P}_1(x_0,\{x_0\}) > 0$).

This verifies all of the conditions of theorem 2 except condition 3.87. It suffices to show that for any $k \in \mathbb{N}$ and any $x \in \mathcal{X}_k$, $P_x(X_m \in A \text{ for some } m \geq 1) >$

0. It has already been shown that for any $x \in \mathcal{X}_1$, $P_x(X_m \in A \text{ for some } m \geq 1) > 0$, since $\mathcal{X}_1$ is the singleton set considered earlier and because this quantity is greater than $P(x_0, A)$, which was positive. Suppose, for induction, that for any $k \leq k_0$ and any $x \in \mathcal{X}_k$, $P_x(X_m \in A \text{ for some } m \geq 1) > 0$. Consider any $x \in \mathcal{X}_{k_0+1}$.

$$P(x, \mathcal{X}_{k_0}) = p_1 P_1(x, \mathcal{X}_{k_0}) \tag{3.92}$$

$$= p_1 \int_{y \in \mathcal{X}_{k_0}} \left[\frac{1}{2}\alpha_1(x,y)\lambda_\uparrow^x(dy) + \frac{1}{2}\alpha_1(x,y)\lambda_\downarrow^x(dy) + h_1(x)\delta_x(dy) \right] \tag{3.93}$$

$$= \frac{1}{2} p_1 \int_{y \in \mathcal{X}_{k_0}} \alpha_1(x,y)\lambda_\downarrow^x(dy) \tag{3.94}$$

$$= \frac{1}{2} p_1 \int_{y \in \mathcal{X}_{k_0}} \alpha_1(x,y)\delta_{g_\downarrow(x)}(dy) \tag{3.95}$$

$$= \frac{1}{2} p_1 \alpha_1(x, g_\downarrow(x)) \tag{3.96}$$

$$= \frac{1}{2} p_1 \min(\phi(g_\downarrow(x))/\phi(x), 1) \tag{3.97}$$

Now, both $\phi(x)$ and $\phi(g_\downarrow(x))$ are positive and finite, so their ratio is as well, and so $P(x, \mathcal{X}_{k_0}) > 0$. This proves that for any $x \in \mathcal{X}_{k_0+1}$,

$$P_x(X_m \in A \text{ for some } m \geq 1) > 0 \tag{3.98}$$

All the conditions are now verified.

The same argument goes through without modification for the simple Markov chain from section 3.7.

In summary the preceding sections have proven the following theorem:

Theorem 3. *Let $(\mathcal{X}, \mathcal{B})$ be the measurable space defined in section 3.3. Let P be the probability transistion function on this space that is defined as a mixture of the component probability transistion functions P_1, through P_5, explained in section 3.8, with mixture weights p_1, through p_5 positive and summing to 1.*

(This Markov chain is a formal version of the Algorithm given in section 3.4.) Then, P satisifies detailed balance with respect to the distribution $\pi'_\circ$ defined by Equation 3.23. Furthermore, P satisfies the conditions of Theorem 2 and therefore, $\pi'_\circ$ is the unique invariant distribution of P. Indeed, there is a set $D \in \mathcal{B}$ such that

$$\pi'_\circ(D) = 1 \quad \textit{and} \quad \sup_{C \in \mathcal{B}} |P^n(x, C) - \pi'_\circ(C)| \to 0 \quad \textit{for each } x \in D. \tag{3.99}$$

Let $f(x)$ be a measurable function on $(\mathcal{X}, \mathcal{B})$ such that $\int \pi'_\circ(dy)|f(y)| < \infty$. Then

$$P_x \left\{ \frac{1}{n} \sum_{j=1}^{n} f(X_j) \to \int \pi'_\circ(dy) f(y) \right\} = 1 \quad \textit{for } [\pi'_\circ]\textit{-almost all } x \tag{3.100}$$

Chapter 4

Examples

This chapter describes the results of a variety of simulation experiments that I have conducted to better understand and evaluate the performance of the posterior mean estimates based on the prior π. In all of these experiments, except where specifically noted, the prior κ on the number of steps K is taken to be $Geometric(\frac{1}{2})$. Other *Geometric* priors are considered in section 4.4 and section 4.7. A few examples in chapter 6 consider *Poisson* priors. There, they are discussed in relation to the convergence theory proven in chapter 5. Most of this chapter, however, concerns evaluating how efficient the $Geometric(\frac{1}{2})$ posterior mean estimate is by comparing it with a wide variety of competing estimation procedures. The first section establishes the standard format for these experiments. It compares the posterior mean estimate with CART and bagged CART estimates and interprets the results. Other methods are considered and compared in section 4.2, namely: a Lasso example that is connected with bagging, three smoothers, and some wavelet-based estimates. Finally, the dyadic Diaconis and Freedman binary regression prior is compared in section 4.3. Section 4.5 takes a step back to analyze the interaction between the data and the model by inspecting how the predictive probability changes as a function of where splits are placed. The final sections experiment with smaller and larger data set size, and also evaluate the performance of the posterior mean on a

more challenging regression problem.

4.1 Comparison with CART and Bagged CART

Since π puts a prior on piecewise constant regression functions, it is natural to ask how its performance compares with conventional estimators that employ piecewise constant approximations. Of particular interest is the *Classification and Regression Tree* (CART) algorithm [4], and the closely related bagged CART algorithm. These methods are briefly reviewed in the next two sections. The impatient reader should skip ahead to subsection 4.1.3, where the methods are compared with the posterior mean estimate on simulated data sets. I have not "filtered" the experimental data at all: these are the originally simulated data sets in their original order. Technically, there has been a certain amount of filtering in the results because I did try using a variety of settings for the CART and bagged CART methods that I do not discuss. The choices that are presented are among the more standard and better performing possibilities. As far as the posterior mean results, these are not filtered at all, except that arguably I would not have a thesis if the results were not interesting.

4.1.1 CART Review

The CART algorithm prescribes how to select a "tree" that represents a good estimate of the unknown regression function (or classification rule). The "tree" terminology connotes the fact that the covariate space $\mathcal{X}$ is recursively partitioned with each piece assigned its own regression value: this recursive partition can be naturally associated (by inclusion) with a graph-theoretic tree. For completeness, a brief description of the CART algorithm is in order. The reader should bear in mind that CART and related algorithms have been in use for many years now and so there are a variety of possible tweaks and alternatives that I do not discuss. The CART algorithm also has important advantages that

the following discussion will not address. Its regression estimate, for example, is unaffected if individual coordinates of the covariate vector are rescaled, so that the estimate does not depend upon the units of measurement. It is capable of dealing with large data sets because of its efficient implementation. It is also very easy to interpret (although this is a risky business since the estimates can sometimes change dramatically with new data).

Suppose for simplicity that the covariate space $\mathcal{X}$ is $\mathbb{R}^d$ (specifically, the case $\mathcal{X} = \mathbb{R}^1$ is of interest at present). In its basic form, the CART algorithm uses coordinate-aligned splits. That is, if a certain subset A of $\mathcal{X}$ is being partitioned, it will be partitioned into the two subsets $\{x \in A \ : \ x_j \leq c\}$ and $\{x \in A \ : \ x_j > c\}$ for some choice of $j \in \{1, \ldots, d\}$ and $c \in \mathbb{R}$. CART proceeds to construct a partition of $\mathcal{X}$ in a greedy manner. That is, it begins by finding the binary partition of $\mathcal{X}$ that maximizes a certain splitting criterion and (proceeding recursively) all subsequent splits are subordinate to this one. Ultimately, the splitting criterion is chosen by the user, and I have chosen to use the ANOVA criterion. To explain this criterion consider that at any given stage of partitioning there is a certain class of possible real-valued functions that are constant on each partition element. Among this class, the function that minimizes the mean of squared residuals to the response data $y_1, \ldots, y_n$ (MSE) is clearly the one whose value on any given element of the partition is the mean of the response values whose covariates "hit" this element. The split that is considered best is the split that results in the greatest possible reduction in this measure of residual error. Having chosen a split, the CART algorithm continues, recursively, to split the resulting subsets. Implicitly, it is building up a "tree" of subsets at each stage with $\mathcal{X}$ forming the root. The recursion terminates whenever there is only a single data point in the current partition element. Call the resulting binary tree of subsets the "full" tree. CART then proceeds to "prune" this tree. This operation depends critically upon a complexity parameter *cp*, which must itself be chosen. Typically *cp* is chosen by cross-validation with the restriction that it not be smaller than

some user specified value. For these experiments I choose *cp* using 10-fold cross-validation and the one-standard-deviation selection rule. This means that if the best achievable cross-validated measure of error (searching over all possible values of *cp*) is *xerr* and the sample standard deviation of *xerr* is *xstd*, the selected value of *cp* will be the largest value whose *xerr* does not exceed *xerr*+*xstd*. To prune the full tree using parameter *cp*, each pair of leaves is considered in turn; if the pair does not improve the splitting criterion by at least the value *cp*, it is removed. Ultimately, every pair of leaves might be removed, resulting in the tree consisting only of the root node. Finally, the pruned tree corresponds to the estimated regression function. It is constant on each element of the (pruned) partition and its value on a given element is the mean of the response values there.

4.1.2 Bagging Review and Discussion

A comparison with the bagging procedure [3] is also relevant. Bagging [3] is a meta-algorithm that can (hypothetically) be applied to any existing classification or regression technique in an effort to improve them. This thesis focuses on bagged CART. Essentially, the bagging idea is just to take many bootstrap resamples of the data set, apply some existing technique to each resampled data set, and then take an average of all of the resulting regression estimates. For completeness, to form a bootstrap resample of a dataset with n items $z_i = (x_i, y_i)$: (1) Independently, choose n integers $N_1, \ldots, N_n$ uniformly at random from 1 to n. (2) Form the new data set: $z_{N_1}, z_{N_2}, \ldots, z_{N_n}$.

It is sometimes stated [43] that bagging is approximately a non-parametric Bayesian procedure, but I think that this is a misleading claim. Bagging, or more specifically bootstrapping, approximates the behavior of a Bayesian who has a (limiting) Dirichlet prior (as in Rubin's Bayesian bootstrap [57]). This is not really a prior for several reasons. Besides the fact that this limiting prior is improper, the "prior" depends on the data. This is not just in a partial

sense (such as the prior I use in chapter 7 in which the "prior" depends on the covariates but not the response) or even in the manner of empirical Bayes procedures which choose some parameters of the "prior" by looking at the data. Indeed, this prior specifies the law of all possible data *relative* to the empirical distribution. This might make sense if the data were multinomial, but for data on a continuous space it is quite problematic. If taken literally it specifies that all future data will consist of elements drawn from the current data set. Notably, for classification, this means that unless the data set happens to contain a head and a tail for each case, then the predictive distribution that the bootstrap prior corresponds to excludes the possibility that the missing flip will ever occur. Indeed, this bootstrap prior can never directly say anything about future datapoints whose covariates are new. Furthermore, there is no meaningful model involved. For example, the prior π implicitly builds in information that says that if a certain region has more heads than tails, then future points in and around this region probably will as well. The bootstrap prior says nothing like that explicitly; if it says that *in effect* it is only because of the fitting method that is forced upon it.

Finally, there is nothing Bayesian about using CART, so how can bagged CART be a non-parametric Bayesian procedure? Why would a Bayesian who believed in a bootstrap prior use CART or neural nets or whatever when he could easily compute his own (very bizarre) posterior and get conclusions directly? Arguably, he might do so in order to get new information to guide his decision because he has observed that CART (say) has worked well on other problems so it probably will work on this one as well. In this sense, bagging could be said to model the behavior of a Bayesian who had a limiting Dirichlet prior (that magically was supported on the data set itself), who then computed the posterior of his "prior," and who then goes to seek the opinion of an "expert." Cleverly, he does so, not only for the dataset actually received, but for a multitude of datasets of size n that are about equally likely under his posterior. In this way, he finds out what CART would think in a variety

of situations that he subjectively considers as possibilities. So far so good. Now he ought to weight these opinions according to how credible they seem in light of the data and his prior opinion about when CART works and when it doesn't. Instead, he now effectively forgets his (Dirichlet) posterior and puts a flat prior on the CART regression results themselves. With his "*newly found prior*," he calculates the decision that minimizes squared error loss, the mean. So, in summary, bagging approximates the behavior of a forgetful Bayesian who looks at the data first, *then* formulates his "prior" and posterior, then *ignores* them, except to ask CART what *it's opinion* would be in the cases that he thinks are likely, then promptly forgets the data altogether as well as any priors or posteriors of his own that he might have held (recently), so he makes up a new uniform *prior* on all the results that he got back from CART; finally he computes the mean according to his latest prior and reports his "findings."

In any case, however dubious as a "Bayesian" procedure, bagging works. One could say that this is because it reduces modeling bias or because it eliminates certain instabilities in CART: both of these arguments make perfect sense to Bayesians and frequentists alike. It is clear, however, that it's not always a good idea, especially when the procedure already has low "instability." In this case, bagging mostly adds noise and reduces the effective size of the dataset somewhat.

I also found one example where bagging was disastrous. Following the bagging procedure strictly, I fed bootstrap resampled data sets into CART, but the result was a mess of indecipherable noise with splits everywhere. This was true even though the CART procedure gave a reasonable estimate on the original data set. Why? Because in the CART step I used cross-validation and this is problematic because CART is working on a bootstrapped sample. Certain repeated data points wind up in both the test and training sets. As a consequence the CART algorithm has less data that it thinks and also thinks that is not over-fitting when it uses a model with too many splits; the pruning procedures became completely ineffective.

Some authors argue that this problem is easy to avoid by simply not bothering to cross-validate within CART, and instead using the full trees. This may be true in some cases, or if one implicitly uses an effective default pruning rule (and not actually full trees), but it failed on my example. To correct the cross validation, some authors "tell" CART which points are repeated so that the whole case is either left in or left out. Instead, for my experiments, I simply used a hand-tuned complexity parameter lower bound. For the disastrous experiment, the lower bound was set to 0; when increased to 0.005 the estimates were still quite rough, but usable. For the experiments that I present in the next section, I set to the lower bound to 0.01, which (admittedly) is the default for the RPART implementation of CART. Still, unless this default is universally good, this leaves a tuning parameter to be set, and I shudder to think about the bizarre computations involved in choosing it by cross-validating bagged CART.

For the experiments involving bagged CART, I used 100 bootstrap resamples. This is a larger number of bootstrap resamples than is generally considered necessary for bagging. By informal Rao-Blackwell-type reasoning one would think that this only serves to reduce Monte Carlo error.

4.1.3 Comparative Simulation Experiment

This experiment involves 10 simulated data sets each containing 1024 data points that were created in the manner explained in the introductory chapter. In this and in future sections these will be referred to as experimental runs 1-10. Briefly, one simply chooses 1024 random uniforms for the x-values, and then flips 1024 independent coins with the success probability of the coin being given by the function f_0 that is indicated by the blue curve in figures 4.1.a, and 4.1.b. These figures also include a red and a green histogram (drawn upside down). These are histograms of the of the x-values for which the coin came out heads or tails respectively. There are 75 bins in each histogram, and to keep

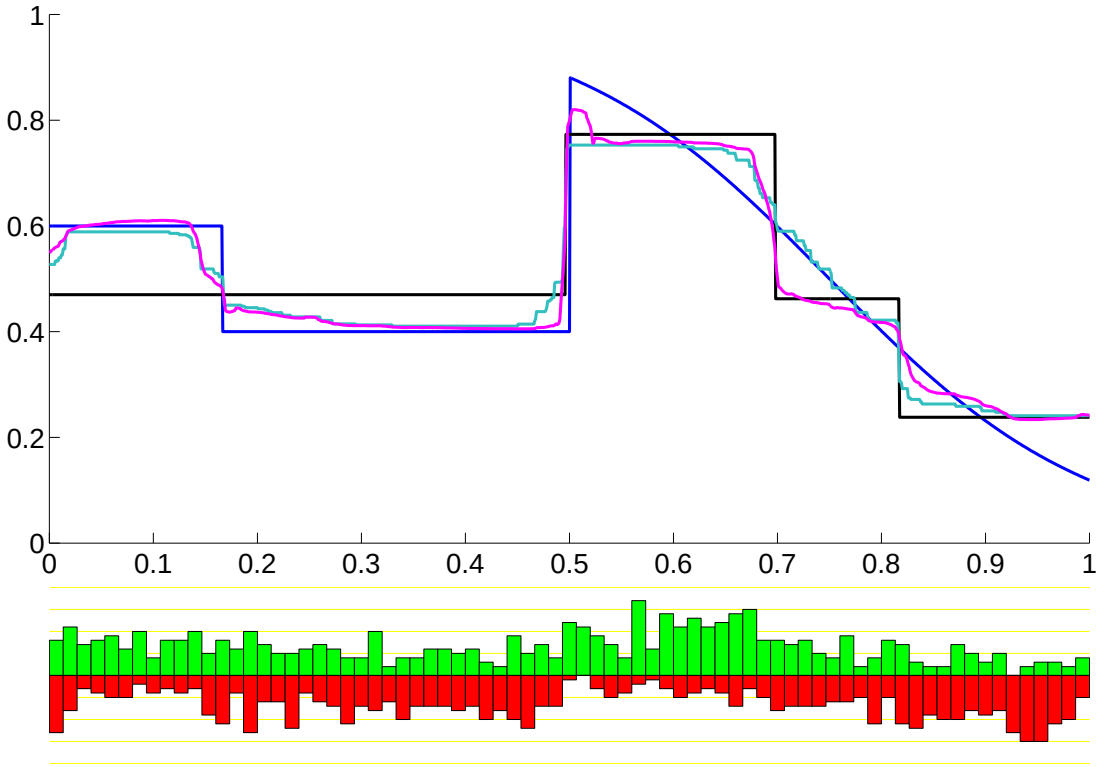

Figure 4.1.a: *Simulation Experiment: Run 1* The CART, Bagged CART, and posterior mean estimates are compared with the true regression curve on a simulated data set ($\alpha = \frac{1}{2}, n = 1024$). CART misses the left hand change-point entirely on this example; bagged CART seems to edge out the posterior mean near 0.75 but they are quite similar over all.
Key: True f (blue), Posterior Mean (magenta), CART (black), Bagged CART (cyan)

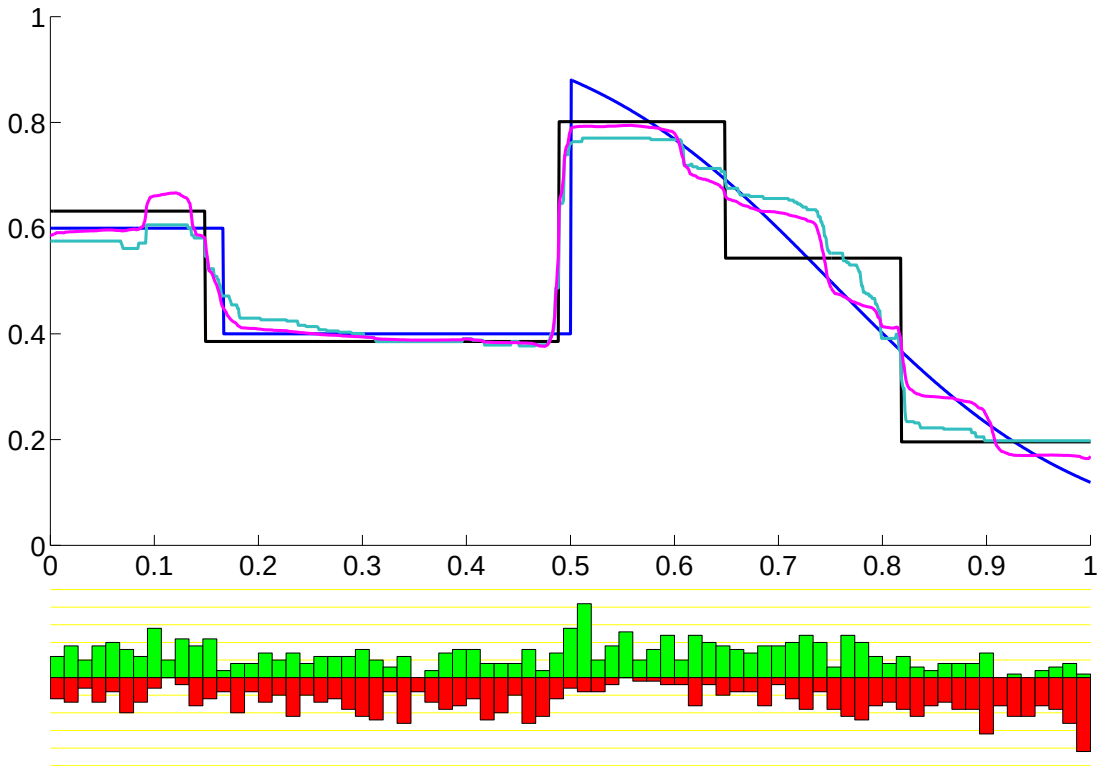

Figure 4.1.b: *Simulation Experiment: Run 2* The CART, Bagged CART, and posterior mean estimates are compared with the true regression curve on a second (equivalently) simulated data set ($\alpha = \frac{1}{2}, n = 1024$). The posterior mean and bagged CART estimates track each other quite closely again.
Key: True f (blue), Posterior Mean (magenta), CART (black), Bagged CART (cyan)

track of the counts, yellow lines are drawn for every 5 events. In these figures, the posterior mean estimate ($\alpha = \frac{1}{2}$) is in magenta, the CART estimate is in black, and the bagged CART estimate is in cyan. The first two experiments are shown in a large format, and the latter 8 are grouped together in Figure 4.1.e.

To be specific, in these experiments to compute the CART and bagged CART estimates, I used the RPART library in S-Plus with default control settings, 10-fold cross validation, the 1-standard-deviation rule to control pruning, and the ANOVA method. I apply the RPART regression algorithms to binary-classification data by assigning the values 1.0 and 0.0 to the two categories. The complexity parameter was restricted to be at least 0.01 (this is the default behavior).

4.1.4 Observations

Here is a summary of notable features in the results of the first two experiments. Some of these features are explained by the subsequent discussion. In experimental run 1 (Figure 4.1.a), the CART estimate only has 4 steps, "leaving out" an important split on the left side. In experimental run 2 (Figure 4.1.b), CART has 5 splits in about the right places. In both figures, of course, CART retains its jagged appearance, but it does do an admirable job of finding the locations where the true curve has change-points. All three estimates share some features with CART, they all make a fairly abrupt change at essentially the same place, somewhere near the location of the true change at $\frac{1}{2}$. They all have substantially more trouble with the change at $\frac{1}{6}$ and this makes sense because it is much easier to detect the difference between two coins with success probabilities 0.8 and 0.4 than between two coins with probabilities 0.6 and 0.4. Surprisingly, at least in experimental run 1, CART seems to be a bit more similar to the posterior mean estimate than to its own bagged version.

Comparing the posterior mean with bagged CART in the first figure, notice that bagged CART smoothes out some steps that the posterior mean leaves

in. Notably this happens near 0.75 where bagged CART comes closer to the truth. For some reason this does not happen near 0.8 where bagged CART is more blocky than the posterior mean. Bagged CART's smoothing was a disadvantage near $\frac{1}{2}$, though. Here the posterior mean makes a much sharper transition and also has an extra "blip" up in the correct direction. Looking at the plateaus, bagged CART has been pulled closer to $\frac{1}{2}$ (away from the truth) than the posterior mean. This is probably due to averaging in a large number of CART trees that omit the left hand split.

Comparing the posterior mean with bagged CART in the second figure, some of the features have remained, but not all. In experimental run 2, the posterior mean takes a much smoother descent on the right than it did before. In this case, the posterior mean is closer to the truth over all. In the first, bagged CART seemed to have an edge. The posterior mean has a blip near 0.1 that bagged CART does not.

4.1.5 Some Explanations

Overall, though, in both experimental runs, the bagged CART estimate and the posterior mean estimate seem quite similar. Consequently, despite the very different way in which the estimates are arrived at, on this example at least the computations achieve a similar result. This is remarkable, especially considering that CART and bagged CART are the result of years of careful problem specific work and tweaking. The posterior mean estimate represent an enormous amount of work too, but most of the work is of a general nature (e.g. MCMC techniques) and not problem specific. Moreover, this prior is very naive (by design) there are many ways to modify it that would improve performance on this example by problem specific tweaking. For example, one could allow both linear regions and constant regions, or impose a suitable (but not too restrictive) dependency among the success probabilities that would help smooth the right side. It may also make sense to space out the locations of the

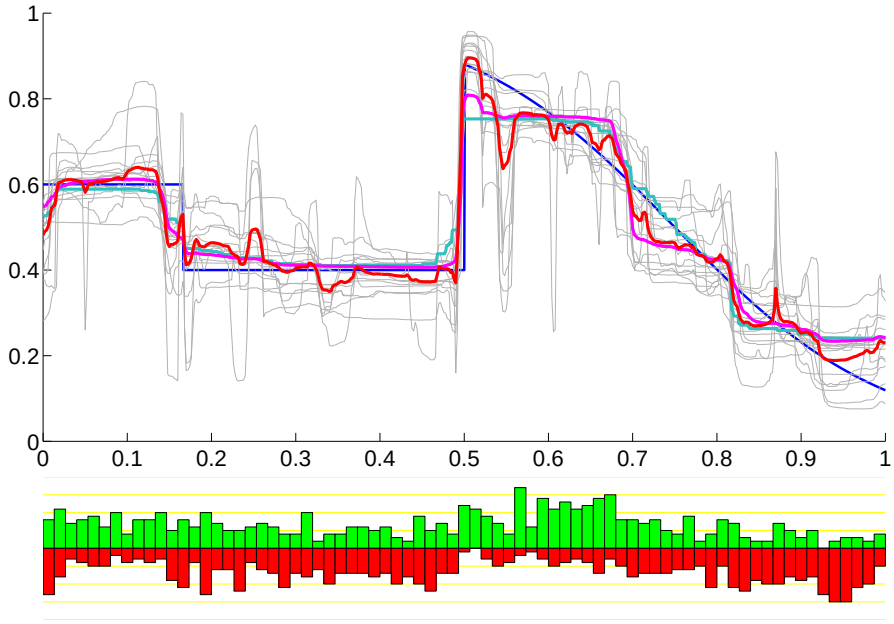

Figure 4.1.c: *The Bagged Posterior Mean Estimate:* Bagging the posterior mean estimate on the data from experimental run 1 produces the red curve which does not seem to change the posterior mean estimate much except to add in some irregular wiggles.
Key: True f (blue), Bagged Posterior Mean (red), Posterior Mean (magenta), Bagged CART (cyan) Posterior Mean on a Fifteen Bootstrap Resamples (gray)

splits explicitly. Rather, this prior is very flat. The choice of *Geometric*($\frac{1}{2}$) as the hierarchy prior is not the result of years of experience, but is, in fact, the first thing I tried. The prior π cannot yet be recommended in general, but that it even performs modestly well "out of the box" on this example is an excellent defense of the Bayesian approach.

Why, is it, though, that the results are so similar? If the results were identical one could hope to prove a theorem about why this was so, but since they are only similar, and since cross-validated CART is not easily amenable to mathematical analysis (much less its bagged variant), I can only speculate that they are similar because they both average together roughly the same functions. They arrive at similar functions because, after all, they use the

same data set. Additionally, under ordinary circumstances there are bound to be similarities between the estimates that CART gives and the estimates that the posterior mean gives (even if I might argue that the posterior mean estimates are preferable). Consequently, if (ignoring decision theoretic discipline), I decided to bag the posterior mean estimate, it "follows" that the *bagged* posterior mean estimate should be close to the bagged CART estimate. Since I believe that the posterior mean estimate is fairly stable under subsampled data (stability under bootstrap resampling is perhaps more questionable, but both questions suggest interesting future research), I conclude that the posterior mean estimate should be close to the bagged CART estimate. If the reader is skeptical that the posterior mean has stability under subsampling, they may be interested in the examples given in section 4.8, these substantiate this claim (but do not address it specifically).

The above argument is merely heuristic, of course, so it is reasonable to ask: what does happen if the posterior mean calculation is bagged? The answer is shown in Figure 4.1.c. The gray curves show the results of computing the posterior mean on fifteen bootstrap-resampled data sets. The red curve is their average: the "bagged posterior mean." Looking at the gray curves, they have many wobbles and spikes, so it seems that the posterior mean is more sensitive to the repeated observations that occur in a bootstrap sample than CART is. As a result, the red bagged posterior mean curve is itself rather wobbly.

4.1.6 Situations in which the Estimates Differ

It is possible to construct examples where the posterior mean estimate would differ more dramatically from the CART and bagged CART estimates. One need only consider circumstances in which CART will reliably perform in a rather special way.

For a first example, recall the CART works by pruning a full tree and that this tree is selected in a greedy manner. The first split that CART chooses,

for example, is usually a very important split, but there are certainly cases where choosing it in a greedy way is suboptimal if one considers the global search for the "best tree." To some extent bagging improves CART's ability to find useful trees because sometimes a resampled data set suggests a different splitting order. Considerations like this are not even an issue for the posterior proper (because it is a theoretical construct), but they are an issue for the actual estimates that get produced by MCMC. Still the issues are different and generally, MCMC methods will perform a much more "global" search over tree space than CART does. This search ability was emphasized as an advantage of Bayesian CART by [6]. In summary, then, in a circumstance where the greedy search has problems, the posterior mean estimate may avoid those problems, and, consequently, give a rather different answer than bagged CART. In my experiments this sort of thing showed up (to a small extent) when I increased the sample size to 8192. This experiment is discussed in section 4.7. The CART estimates preferred a split in the middle of the right-hand slope that the posterior mean avoided.

A second example occurs if the x-axis is transformed. One-dimensional CART estimates are invariant with respect to this, while the posterior mean is not. In some senses this property is desirable; it seems "scientific." It is not always desirable though: suppose there is a large amount of data from 0 to 0.1 and from 0.9 to 1 but no data in between. Roughly, CART will treat this in the same fashion as if there were no gap; essentially it only looks at the ordered values. If it splits the gap at all, it will split in exactly the middle of the gap: between the rightmost of the left-hand data and the leftmost of the right-hand data (ignore the fact that this is not *strictly* invariant under transformations). This effect does not go away under bagging (although it might get smoothed a bit as the endpoints of the gap change).

However, the posterior mean estimates will be quite different. Notably, it will make a smooth transition from the regression value on the left to the value on the right. Additionally, because this gap is especially large and the prior

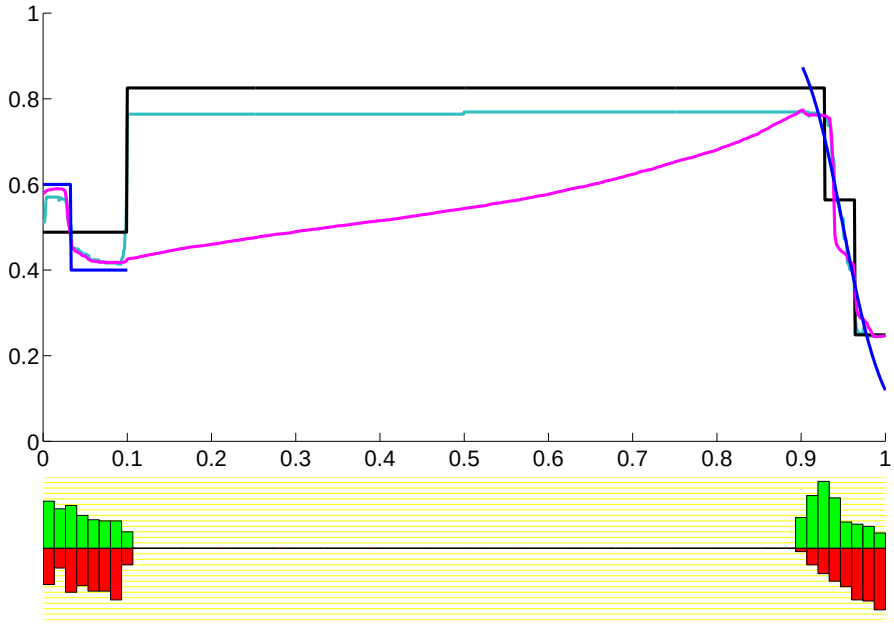

Figure 4.1.d: *Transformed Data with a Gap:* For this example, the data from experimental run 1 was transformed nonlinearly to create a gap. Notice the smooth transition that the posterior mean makes along the gap.

specifies that split points are put down uniformly, it is quite likely that the gap will be split at least once or twice. If the gap is split twice or more, then the middle intervals will include no data at all. When there is no data (or even little data), the prior kicks in to specify that it thinks that the success probability is uniformly distributed from 0 to 1 and consequently that the mean value of the success probability on this middle interval is $\frac{1}{2}$. Because cases like this get averaged in, the posterior mean should show some shrinkage to $\frac{1}{2}$ on the gap. Shrinking to $\frac{1}{2}$ is, of course, not always ideal, but in such a case one ought simple to modify the prior and/or loss function. To construct an example with a gap, I took the data from experimental run 1 and transformed it so that the left half of the data now lies in $[0, 0.1]$ and the right half lies in $[0.9, 1]$. The result of computing the posterior mean is shown in Figure 4.1.d

These features make obvious intuitive sense. Furthermore, if a 95% subjective confidence interval were formed by asking at each point $x \in [0, 1]$ for the

smallest interval containing 95% of the posterior mass, it would grow wider in the middle. In contrast, if naive confidence bands were formed around the CART or bagged CART estimates (by using bootstrapping perhaps) they would make rather little sense.

4.1.7 Posterior Mean Behavior

In the first experimental run, CART left out an "obvious" split, while in the second it put it in. Sometimes CART will also "add in" splits that it should not have; this is very sensitive to the particular data set and the pruning rules that are used. Bagged CART averages all of these together (in some special way that is hard to formulate, except algorithmically) to arrive at its smoother curve.

In contrast, I imagine that the posterior mean is considering each of these possibilities and giving them an appropriate weight before averaging, in order to give its best estimate. Examining the posterior mean curve closely, one can see that wherever CART takes a step, the posterior mean also moves more abruptly than normal. Both estimates are, after all, both looking for steps, and the locations that CART chooses are bound to be special parts of the data set, often containing a run of heads on one side and a run of tails on the other. Surely this feature would stand out to both methods.

The reverse is not true, however. Consider the "bump" in the posterior mean, just to the right of $\frac{1}{2}$. This results because the posterior considered functions with additional splits in this area and gave them weight. It is not, after all, the result of averaging a large quantity of individually pruned trees, but the result of averaging over all trees (in principle) with appropriate weights. CART trees, if allowed to have extra splits would have included this one as well. Along this line, a modest improvement to the bagged CART procedure might result if in addition to using different bootstrapped datasets, one sometimes used different pruning criteria as well, and then averaged the less penalized

trees in together with the more penalized ones (with appropriate weights).

As an aside, the posterior sometimes becomes more "sure" about the location of a split than it "really ought to." This happens because of the mismatch between the truth and the prior (or perhaps more accurately, because of the mismatch is between π and my own subjective prior). The posterior is doing the absolutely optimal thing if the prior π is true (indeed, it is an admissible estimator and there is no easy way to tell if the others are or not), but according the prior, functions like the one pictured in blue with a smooth transition are impossibly rare. When faced with data that could have resulted from a smooth transition, but might also credibly be created by a step function with two steps, the posterior only considers the latter possibility. This is especially apparent when there is a run of heads and then a run of tails occurs by chance (as it is bound to do from time to time). The posterior will concentrate more tightly around this cut-point than makes sense if one considers a smooth transition to be a credible alternative explanation. The posterior does, of course, allow for the possibility that there are multiple splits, but if the success probabilities on each side are reasonably similar, there may not be enough data to make this possibility stand out and the single split will remain the most prominent feature of the posterior mean. This results in a stair-step appearance that does not go away with larger sample sizes (see section 4.8), although the steps tend to get smaller. Indeed, it appears that the stair-step shape grows more prominent for larger n. Perhaps this is because a larger data set also is more likely to have at least a few very long runs.

4.1.8 Experimental Runs 3-10 and a Summary

The results from experimental runs 3 through 10 are shown in Figure 4.1.e. By and large, the same observations made before above apply to these examples. Bagged CART and the posterior mean track each other quite closely although each occasionally takes a "wobble" that the other does not. Broadly speaking

both estimators still retain visible traces of the CART-type functions that they are averaging together. In particular, both usually have some "stepiness" in their appearance; the averaging "softens" this, but does not eliminate it. This is especially visible in experimental run 10 in which both estimates also follow the CART estimate quite closely. CART leaves out the left hand split on 5 of the 10 experiments.

	CART	Posterior Mean	Bagged CART
1	0.0891	0.0522	0.0527
2	0.0776	0.0478	0.0568
3	0.0893	0.0587	0.0710
4	0.0996	0.0660	0.0737
5	0.1101	0.0659	0.0683
6	0.1130	0.0609	0.0799
7	0.0779	0.0604	0.0593
8	0.0907	0.0669	0.0617
9	0.1009	0.0720	0.0718
10	0.0671	0.0531	0.0587
$\widehat{\mu}$	0.0915	0.0604	0.0654
$\widehat{\sigma}$	0.0147	0.0076	0.0088

Table 4.1: Numerical Summary of $\mathcal{L}^2$-norm errors on the ten experimental runs

A numerical summary of these ten experiments can be made by computing the $\mathcal{L}^2$-norm of the error between the estimated curve and the truth. This summary is given in table 4.1.8 and illustrated by the scatter-plot in Figure 4.1.f. The black points compare CART to the posterior mean. The cyan points compare bagged CART to the posterior mean. In each case, the x-axis is the $\mathcal{L}^2$-norm error of the posterior mean, and the y-axis is that of the competitor. Obviously, small numbers are preferred, and because the black points lie exclusively above the identity line on these ten experiments, the posterior mean is preferable here. The performance of the bagged CART and the posterior mean estimates is much closer, although the posterior mean's performance is slightly better on average.

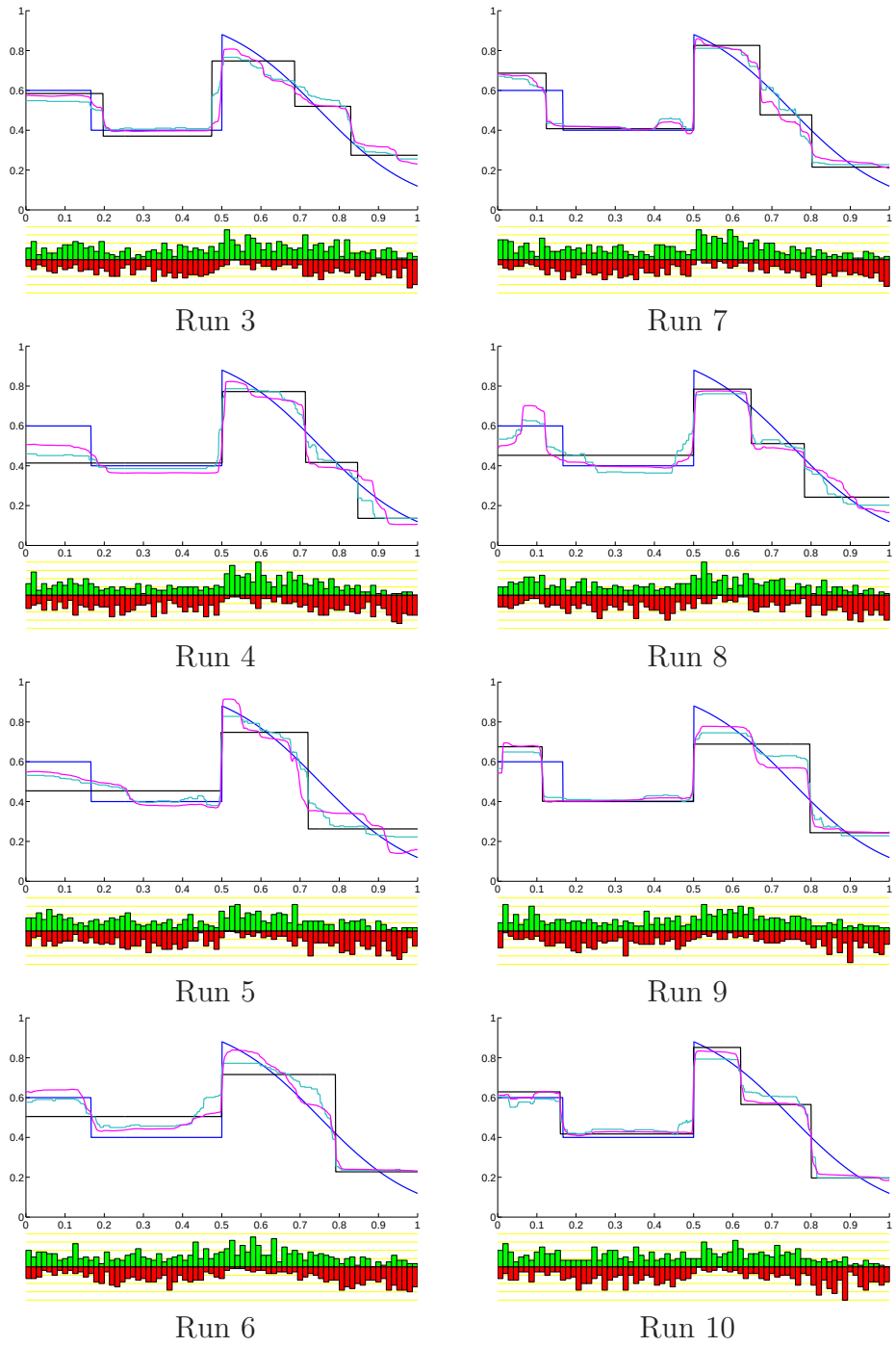

Figure 4.1.e: *Simulation Experiment: Runs 3-10*
Key: True f (blue), Posterior Mean (magenta), CART (black), Bagged CART (cyan)

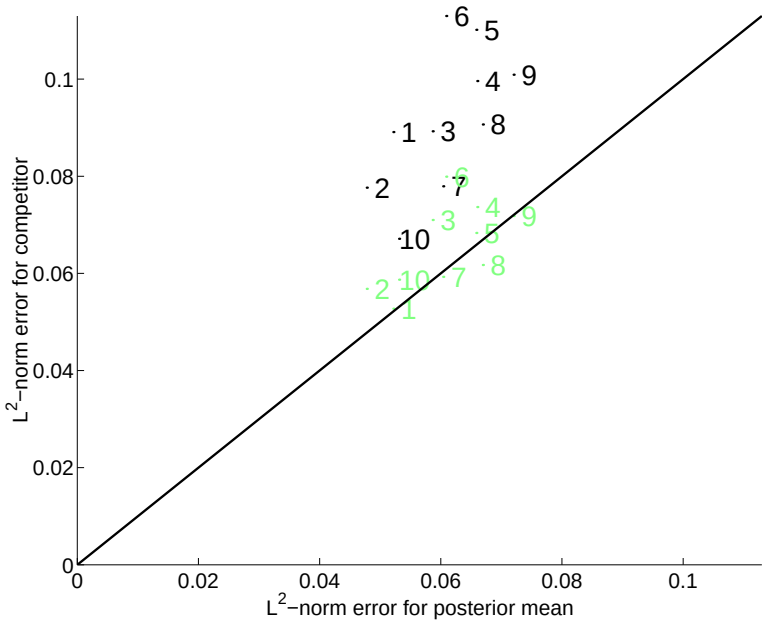

Figure 4.1.f: *Comparative Scatterplot:* The results of the ten experimental runs are summarized by $\mathcal{L}^2$-norm error (also known as $\sqrt{MSE}$). On the x-axis is the error committed by the posterior mean estimate. On the y-axis is the error committed by the CART (black) or bagged CART (cyan) estimates respectively.

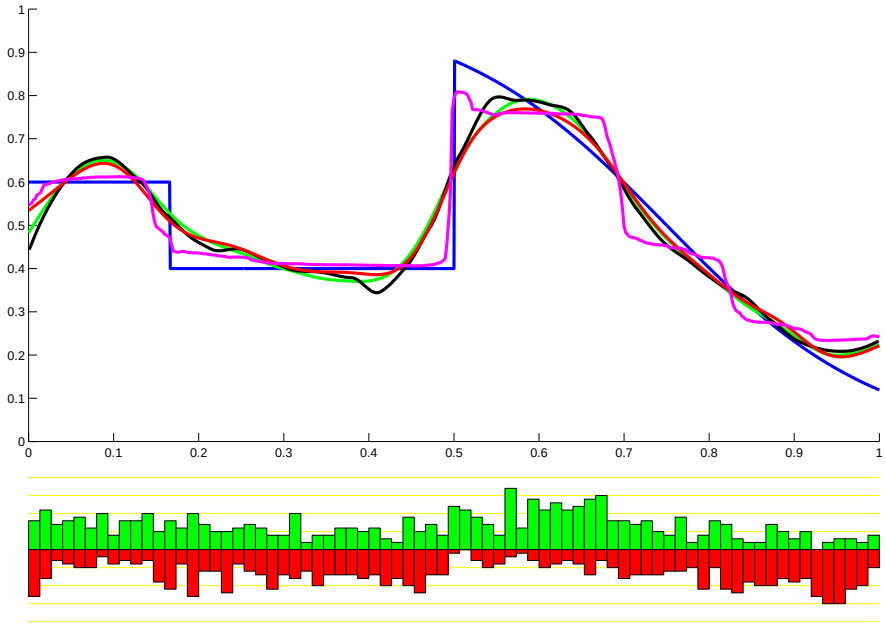

Figure 4.2.a: Three Smoothers:
Key: True f (blue), Posterior Mean (magenta), Loess (black), Smoothing Spline (green), Kernel Smoother (red)

4.2 Comparison with Other Popular Methods

This section compares the posterior mean estimate with a variety popular techniques on the data from experimental run 1. Throughout this section, as usual, the true curve is plotted in blue and the posterior mean with $Geometric(\frac{1}{2})$ prior is plotted in magenta.

4.2.1 Smoothers

Figure 4.2.a shows the results from running three standard smoothers. The results are little surprise. The three smoother's estimates are quite similar over all on this example. They over-smooth the jumps, but partially make up for this by giving a smoother approximation on the smooth half. They also all take a turn at the ends; this is consistent with the data which happens to

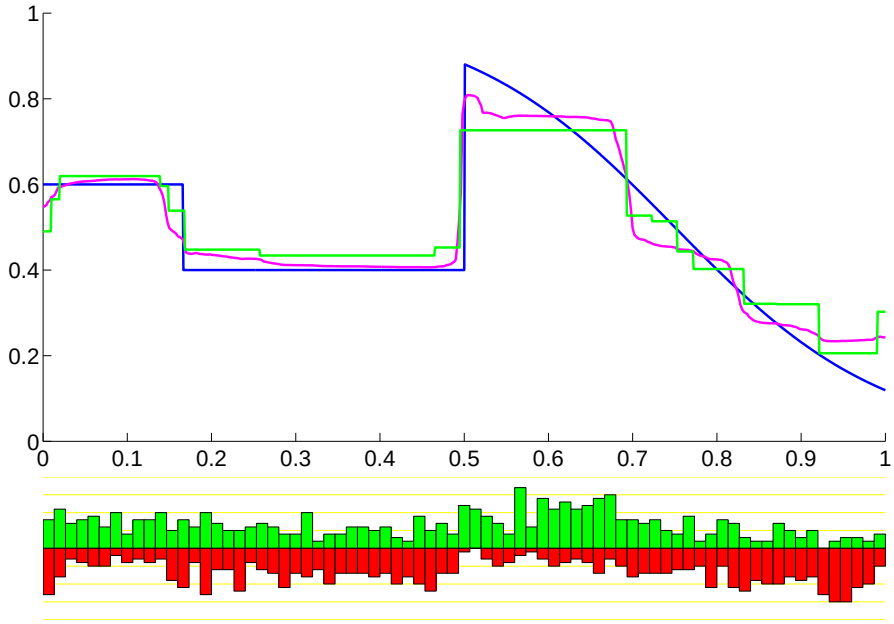

Figure 4.2.b: A Lasso Estimate:
Key: True f (blue), Posterior Mean (magenta), Lasso (Cp) (green)

behave somewhat unusually there.

Loess (plotted in black) fits a locally-weighted linear regression at each point to make its estimate. Smoothing splines (in green) use efficient computational tricks to compute the regression curve that optimizes a tradeoff between small MSE and small integrated second derivative. Gaussian kernel smoothers (plotted in red) take a weighted average of response values near a point to predict. For each method, I chose a smoothing parameter that seemed to give results that were about as good as possible.

4.2.2 LARS/Lasso/Boosting

The green curve in figure 4.2.b, shows the result of using a Lasso penalized regression. This was particularly easy to do using using the Least Angle Regression (LARS) software [30]. Like any linear regression, the results depend

on what basis is chosen. For this example, the basis is constructed by considering the function $\mathbf{1}_x \geq c$ for 1200 values of c evenly spaced from 0 to 1. To encode a datapoint with covariate $x \in [0, 1]$ under this basis, evaluate the 1200 functions at x and pack the results into a vector: this vector becomes the covariate that the regression uses. Finally, Lasso regression resembles ordinary regression except for one critical difference: the regression parameter β is penalized by $\lambda||\beta||_1$, where λ is a tuning parameter. For this example, the parameter λ was chosen using a Cp criterion.

Constructed in this way, the Lasso regression estimate should be quite similar to the estimates that would be arrived at using other important techniques such as "Boosting stumps" and the least angle regression method. The similarity between these different methods is discussed in [30].

As can be seen from the figure, the Lasso estimate is has some appealing features. It is piecewise constant, but it also takes a fairly large number of steps and spaces them out usefully along the smooth transition on the right side of the figure. It does a reasonable jump of "detecting" the two change-points. On the other hand, it does not go as low as it should from $\frac{1}{6}$ to $\frac{1}{2}$, nor as high as it should to the right of $\frac{1}{2}$.

4.2.3 Wavelets

Figure 4.2.c compares some estimates that were based on wavelet techniques. The red, black, and dotted curves show various wavelet reconstructions of the regression curve. Overall, I think the results are quite disappointing; artifacts from the particular basis used show through clearly into the estimate. Since the data are not regularly spaced, some accommodation is necessary to use conventional software (e.g. Wavelab). Algorithms exist that apply directly to irregularly spaced data, but I did not successfully locate any working implementations. Instead, the dotted curve shows the wavelet reconstruction that results from simply using the ordered covariate values as if they were regularly

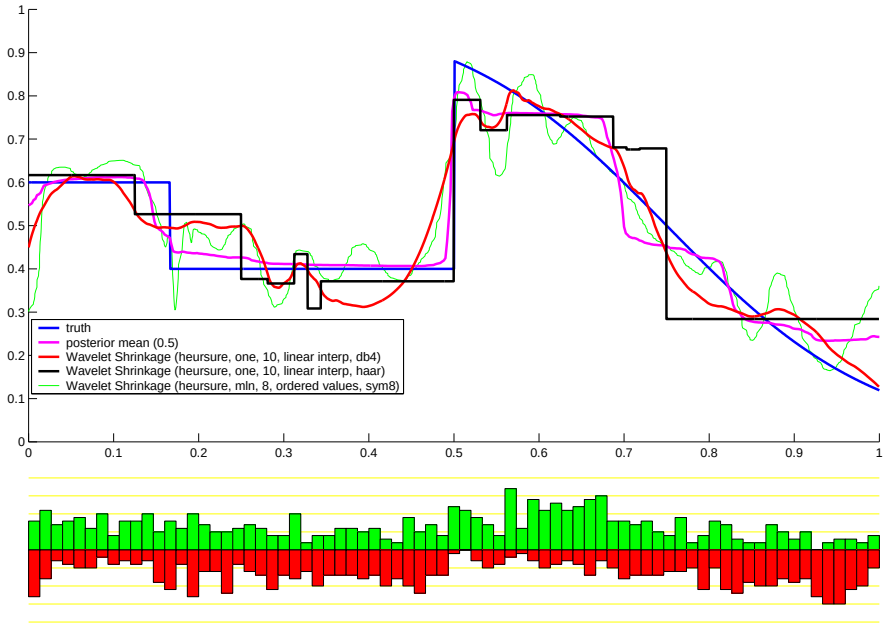

Figure 4.2.c: Wavelet Estimates:
Key: True f (blue), Posterior Mean (magenta), Wavelet 1 (green), Wavelet 2 (red), Wavelet 3 (black)

spaced and then extrapolating back to the irregularly spaced reality.

There are a number of problems with this approach and some recent work has developed more sophisticated schemes. For the red and black curves, I tried one of the simplest [14] which recommends using direct linear interpolation to produce values on a fine grid and then applying wavelet shrinkage methods to this gridded data. The estimates were computed by the `wden` function in Matlab. It has a large number of options, most of which (quite frankly) I do not understand. These include the choice of a threshold selection rule from among four options, the choice to use hard or soft thresholding, an option titled "multiplicative threshold scaling" with three options, the choice of the level at which to compute the coefficients, and finally the name of the wavelet family to use (many options). For someone with as little experience with wavelet methods as me, these options are not a feature but a drawback. Furthermore, experimenting with the different options, they all seemed to make a difference. A reasonable, but not heroic effort was made to choose working parameter values. In any case, for the illustrated curves, the Matlab commands that were used are:

```
Xi=seq(min(X),max(X),2^12); % a fine grid of X-values
Yi=interp1(X,Y,Xi);         % on which to interpolate the response
Yhat1 = wden(Y, 'heursure','s','mln', 8,'sym8'); % 1: green
Yhat2 = wden(Yi,'heursure','s','one',10,'db4' ); % 2: red
Yhat3 = wden(Yi,'heursure','s','one',10,'haar'); % 3: black
```

4.3 Comparison with Dyadic Prior

For comparison, Figure 4.3.a shows the result of computing the posterior mean resulting from the Diaconis and Freedman dyadic binary regression prior [25]. Like the prior π studied in this thesis, this prior chooses a random partition and assigns independent success probabilities to each partition element. However,

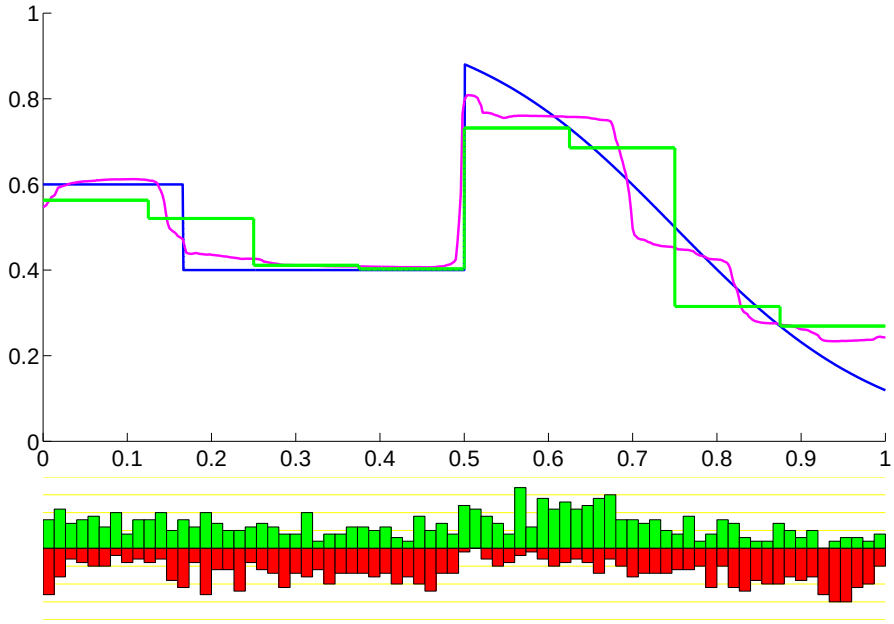

Figure 4.3.a: *Dyadic Posterior:* The posterior mean of the Diaconis and Freedman dyadic prior on the data from experimental run 1 is drawn in green

this prior uses a dyadic partition, splitting the data into 2^k equal pieces for some $k \geq 0$. The consistency of the resulting estimates is guaranteed for any choice of hierarchy prior, except perhaps when the true regression function is identically $\frac{1}{2}$ and the data is pure noise. For this case, certain priors will be consistent and others will be inconsistent. It was of interest, then to try one on the boundary. For this reason the prior on hierarchy level K that was used assigns: $P(K = k) = (1 - \beta)\beta^k$ for $\beta = \exp(-2^{-\frac{1}{4}}) \approx 0.431$. In fact, for this data, the results were stable over a wide range of choices of β.

Since the prior is dyadic, it has no trouble at all nailing the split at $\frac{1}{2}$; of course, it does not have such good luck for the split at $\frac{1}{6}$. The posterior strongly favors a model with around 8 steps.

4.4 Dependence on the Parameter of the Geometric Prior

In this section, consider a departure from the *Geometric*($\frac{1}{2}$) hierarchy prior. Figure 4.4.a shows what the posterior mean estimate would be for the data from experimental run 1, if a *Geometric*($1-\alpha$) prior is used on the number of steps K. As usual, the true response curve is indicated in blue, and the posterior mean estimate for $\alpha = \frac{1}{2}$ is drawn in magenta; the posterior means for other values of α are also drawn. As can be seen in the figure, as α ranges from small values (short tail prior, dotted black curve) to large values (long tail prior, solid black curve) there is not so much difference in the posterior mean estimate except that certain small bumps and wiggles that are suppressed for the small α values become visible for the larger values. Indirectly, this experiment also provides a check on the stability of the Monte Carlo estimates of the posterior mean; it is unlikely that there would be such close agreement among these independently computed estimates if the MCMC was not working reasonably well. To make a more detailed comparison, it would be sensible to pool the sampled regression curves and use an importance sampling technique to compute combined results. I do not pursue this here.

It is also of interest to know how many splits the posterior is using. Figure 4.4.b shows the posterior on the number of steps K for three *Geometric*(α) hierarchy priors: $\alpha = 0.1$, 0.5, and 0.9 respectively. The difference between the priors has a more pronounced effect here. Under a *Geometric*($1-\alpha$) prior with $\alpha = 0.1$, complex models are quite rare and consequently, the posterior on step size shown in the top panel has a fairly short tail. Notice, though, that even for this conservative model, the likelihood has been able to overwhelm the conservative prior enough to shift most of the posterior mass onto models with 5 splits.

For the bottom panel, $\alpha = 0.99$, which corresponds to a rather slowly decaying tail. Notice, though, that the tail of the posterior is not nearly this

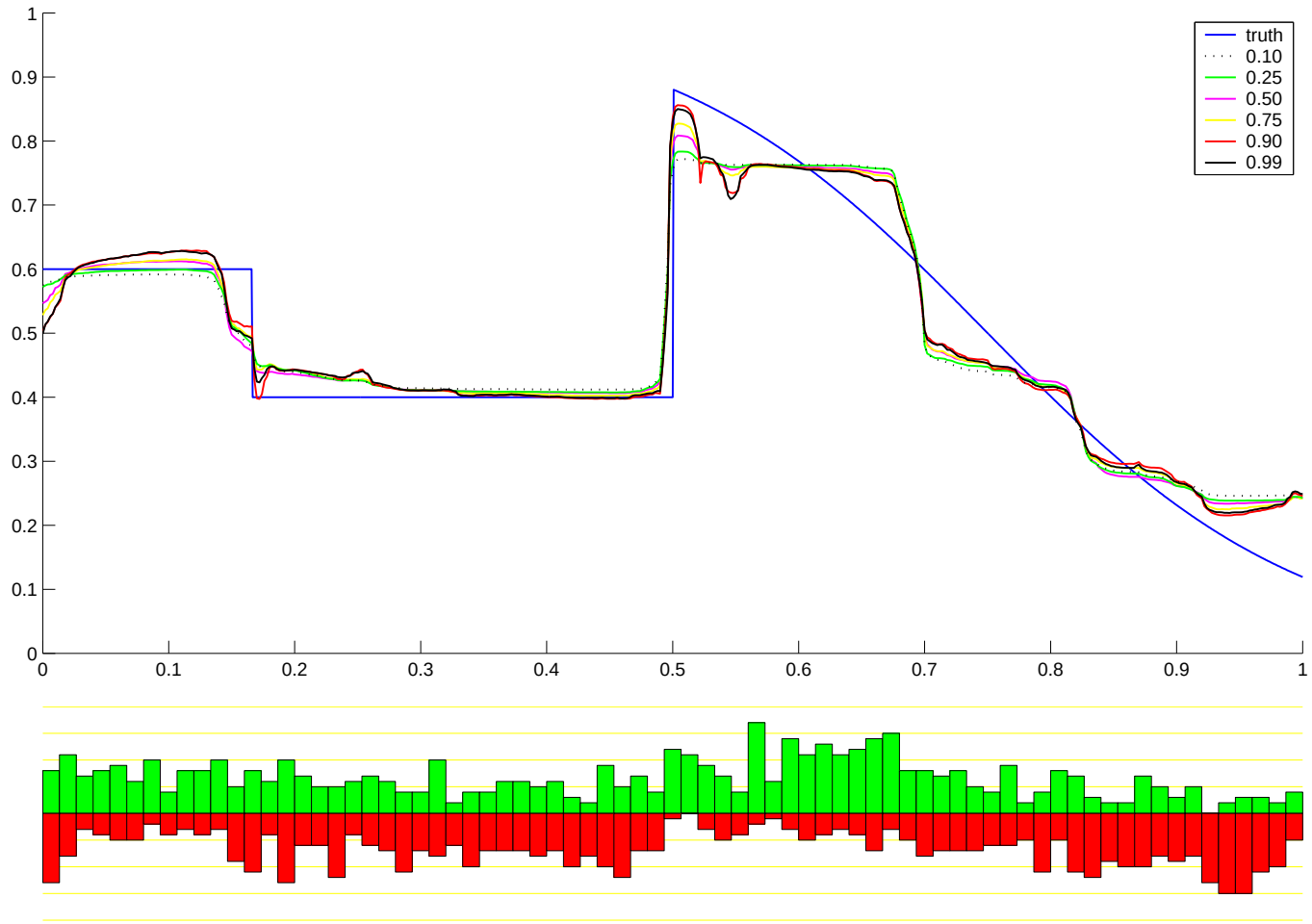

Figure 4.4.a: *The Posterior Mean for a variety of Geometric Priors*: Consider the data of experiment 1 and compute the posterior under a $Geometric(1-\alpha)$ prior.

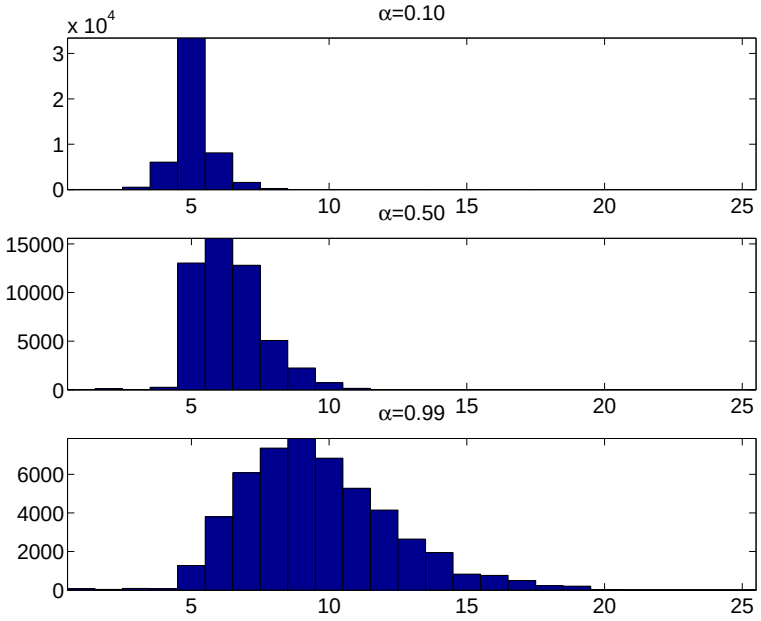

Figure 4.4.b: *The Posterior on Model Size* Consider the data of experiment 1 and employ a *Geometric*$(1-\alpha)$ prior. A histogram of the number of sampled regression functions that had k steps is shown for three values of α: 0.10 (top), 0.50 (middle), 0.99 (bottom)

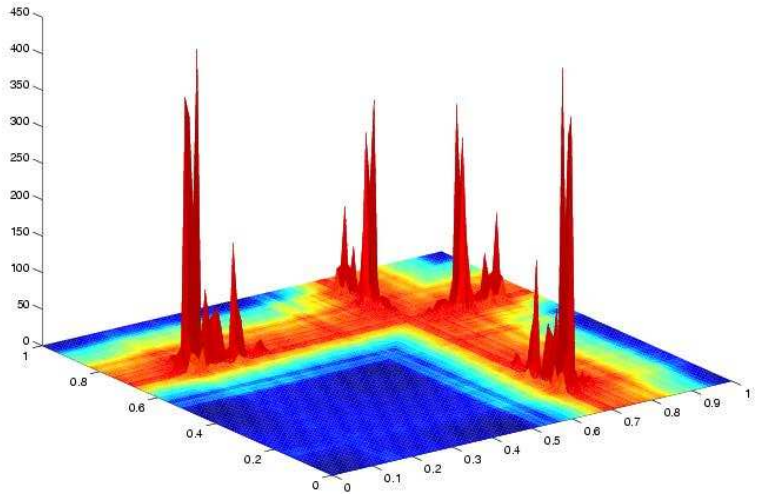

Figure 4.5.a: *A Marginal Likelihood Surface*: A slice of the posterior for the 3 change-point case ($k = 4$) is shown. One change-point is fixed at 0.491 (not shown) and the location of the other two vary along the x and y axes of plot respectively.

long: models with a large number of steps k *can* fit the data well, but also have a large number of parameters: $2k - 1$. This tends to down-weight them as a group. When the posterior is marginalized to yield the posterior distribution on the number of steps, an account is taken of "how many" of these more complicated models give a good fit as well as how good the fit itself is. Because of this trade of, models with 9 steps are the most common.

The middle panel, corresponds to the *Geometric*($\frac{1}{2}$) prior that has been the subject of so many experiments. Notice that models with fewer than 5 steps have almost no effect on the posterior mean; most of the mass is on models with 5 to 9 steps: 6 is the most common choice.

4.5 The Predictive Probability Surface

To better understand the interaction of the data and the posterior on experimental run 1, reference to Figure 4.5.a is useful. It demonstrates just how spiky, multi-modal, and (in particular) non-normal the posterior's density can be. It shows a slice of the posterior for the 3 change-point case (i.e. four steps). One change-point is fixed at 0.491 (not shown) because this was the most likely location for a single split; this location accounts for the jump at $\frac{1}{2}$ in f_0. The other two change-points are allowed to range over the x and y axes of the plot. More technically, what is plotted is a self-normalized version of the function $\phi(x)$, defined by Equation 3.23 where $x = \big(4, (0.491, \textit{x-coord}, \textit{y-coord})\big)$. Essentially $\phi(x)$ computes the likelihood that the splits occur at a given location; it marginalizes out the different possible choices for the success probabilities. The height of the surface follows $\phi(x)$, and the color follows $\log(\phi(x))$, so that the small-probability structure is also visible. The x and y axes are symmetric, of course, because splitting at a and b is the same as splitting at b and a. Similarly the function is largest along horizontal and vertical "bands." This is because when one split is in a particularly fortuitous place, it tends to improve the fit over all, even if the placement of the second split is suboptimal. The highest two peaks (near the opposite corners), represent splitting on the left (in the vicinity of $\frac{1}{6}$) to take care of the jump that is there, and on the right (in the vicinity of 0.75) to split the smooth transition region into its higher and lower halves. The secondary peaks near the far corner, represent splitting the smooth transition in two places and ignoring the left half (recall that this was the choice made by CART on this data).

4.6 Behavior on a Small Data Set

It is interesting to see how the posterior responds to individual data points; this is most easily seen in a very small dataset. In Figure 4.6.a I consider a

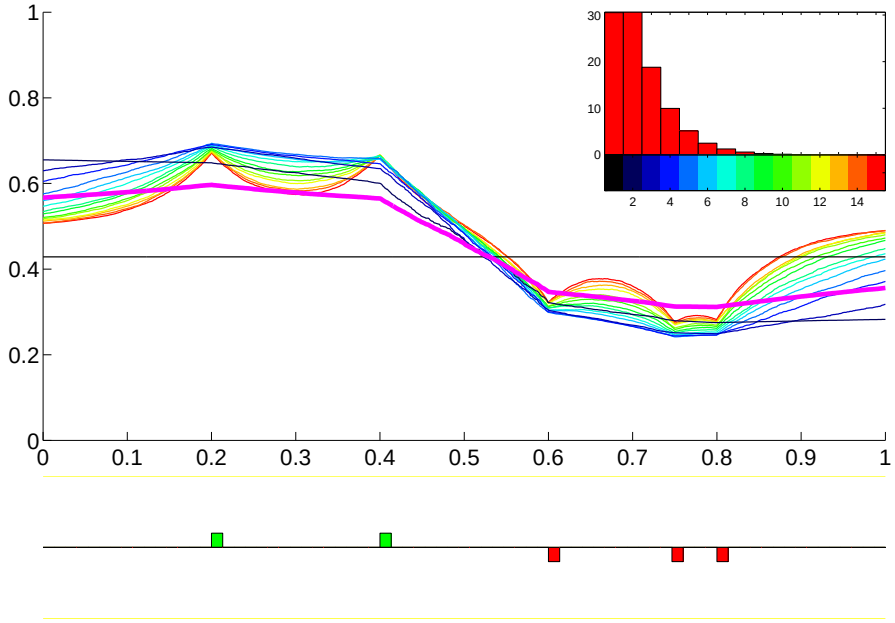

Figure 4.6.a: *Small Data Set Experiment:* On this five element data set, the posterior mean curve in magenta is the weighted (c.f. inset histogram) composition of the colored curves. These curves represent the mean contribution from models with a fixed number of steps which ranges from 1 to 15

data set with 5 data points, that, as usual are shown by the two histograms. In this case there are two heads on the left and three tails on the right. The posterior mean (magenta) is seen to respond in a smooth way, except at the data points where it remains continuous but takes a (small) sharp turn. The other colored curves represent the posterior mean when the number of steps that the function is allowed to take is fixed a-priori. The flat black line, for example, is the posterior mean when no splits are allowed (one step). The red curve, on the other hand is the posterior mean if 14 splits (15 steps) are required. Notice how the red curve, especially, drifts back towards $\frac{1}{2}$ for x-values that are not close to the data points. This happens because with 14 splits and 5 data points, there are bound to be many empty intervals and those, necessarily, fall back on the prior. The posterior mean curve for a $Geometric(\frac{1}{2})$ prior on the number of steps is the weighted average of these curves, where the weights (as percentages) have been tabulated by the inset histogram. Since the 1-split model fits this data so especially well, it winds up contributing about as much as the contant model (which cannot be ruled out with so little data) to the final result. The contribution of the higher models is not forgotten, though; notice how the posterior mean (magenta) drifts back to $\frac{1}{2}$ slightly on the left of 0.2 and the right of 0.8. Overall the posterior mean is conservative; it does not, for example, split the data in half at 0.5 and declare the left hand mean to be 1 and the right hand mean to be 0.

4.7 Behavior on a Large Data Set

The data from experimental run 1, consisted of 1024 (x, y) pairs with the x-values drawn uniformly from $[0, 1]$ and the y-values drawn $Bernoulli(f_0(x))$. This section answers the question: how does the posterior change if this data set is enlarged to have 8192 datapoints by generating additional data from this model? As usual the true curve is drawn in blue and the posterior mean for the $Geometric(\frac{1}{2})$ prior is drawn in magenta. The data set, as is visible from the

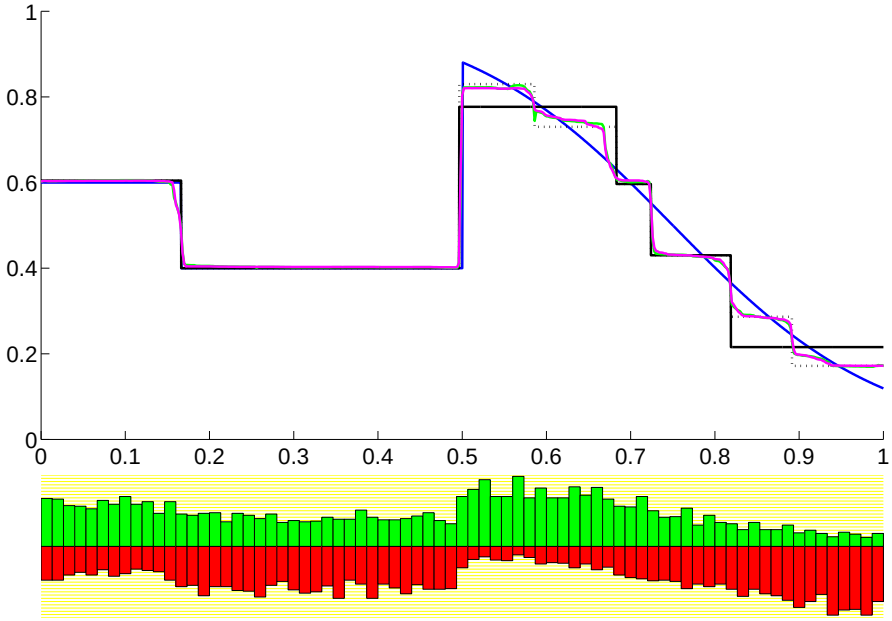

Figure 4.7.a: *Large Data Set Experiment:* The data from experimental run 1 is extended to a large data set of size 8192
Key: True f (blue), Posterior Mean ($\alpha = \frac{1}{2}$) (magenta), Posterior Mean ($\alpha = 0.995$) (green), CART (black), min-*xerr* CART (dashed black)

histograms, is getting rather large. Pleasantly, the posterior mean estimate is also giving an accurate estimate of the true curve. Somewhat disappointingly, though, the step functions on which the prior is based have not gone away. Although they are smaller, they are clearly visible in the magenta curve.

The CART estimate for this data is shown in black. A couple of observations need to be made. First of all, because the 1-standard-deviation pruning rule was used, the CART curve only has 6 splits. Looking through the full tree, the model that minimizes the cross-validated error has two additional splits and is drawn by the dotted black line. This model agrees quite closely with the posterior mean estimate, but two of its splits were pruned away when the 1-standard-deviation pruning rule was used because the standard deviation of the *xerr* is not small enough. It might be selected automatically if a more intensive cross-validation were used. Finally, note that both CART curves minor artifacts (when compared to the posterior mean) that result from the greedy nature of the full tree.

One might suppose that with this much data, it ought to be possible to fit a model with many more steps. This does not seem to be true (at least for a prior that models the success probabilities *independently*). For example, if the full CART tree is manually pruned to have only one or two additional splits beyond the 8 that were used by the minimum *xerr* model, the additional splits visibly degrade the fit. To understand this, consider that the right half of the data should contain around 4096 points. By chance, they will not be (quite) evenly distributed over this half but for simplicity suppose that these "4096" points are divided evenly into the 6 intervals selected by the larger CART model. This leaves around 680 points in each partition cell. Recall that the variance of a $Binomial(n, p)$ random variable is $np(1-p)$, so that the standard deviation of the estimated success probability on each of these partition cells is going to be around 0.02. Considering that the regression estimator has to not only detect a difference, but also locate a good choice of split, and optimize the accuracy of the estimated success probability, it does not seem too unreasonable that the

average jump in success probability between neighboring cells is around 0.1.

Also shown (by the green curve) is the result of computing the posterior mean when a $Geometric(1-\alpha)$ prior is used for $\alpha = 0.995$. Interestingly, this long tail makes little difference and the green curve barely peaks out from under the magenta one.

4.8 The Effect of Sample Size

The previous section developed a extended version of experimental run 1 that contained 8192 data points. In figure 4.8.a, this data is analyzed in more detail. Smaller data sets are formed by taking the first n data points for n ranging from 64 to 8192 by powers of two. It is very pleasant to see how the posterior mean incrementally grows closer to the truth. At first, the steps on the left are almost ignored (with so little data, any pattern they contain could have resulted from noise), but gradually they fill in. The transitions near the change-points in f_0 become very sharp and the estimates of the smooth transition steadily improve.

4.9 The Effect of Sample Size: a Harder Example

For the final example, I consider a much harder regression function. It is depicted in blue in Figure 4.9.a. It was formed by taking multiple copies of the original f_0 and shrinking them to half their size repeatedly. A data set with 8192 point is simulated and the posterior mean is calculated for subsets of increasing size. As had been hoped, the features get filled in as the data size increases incrementally. The larger features rise above the noise first and then the smaller, so that for this regression function, the approximation seems to grow better as n increases on the right first, but then steadily spreads to the

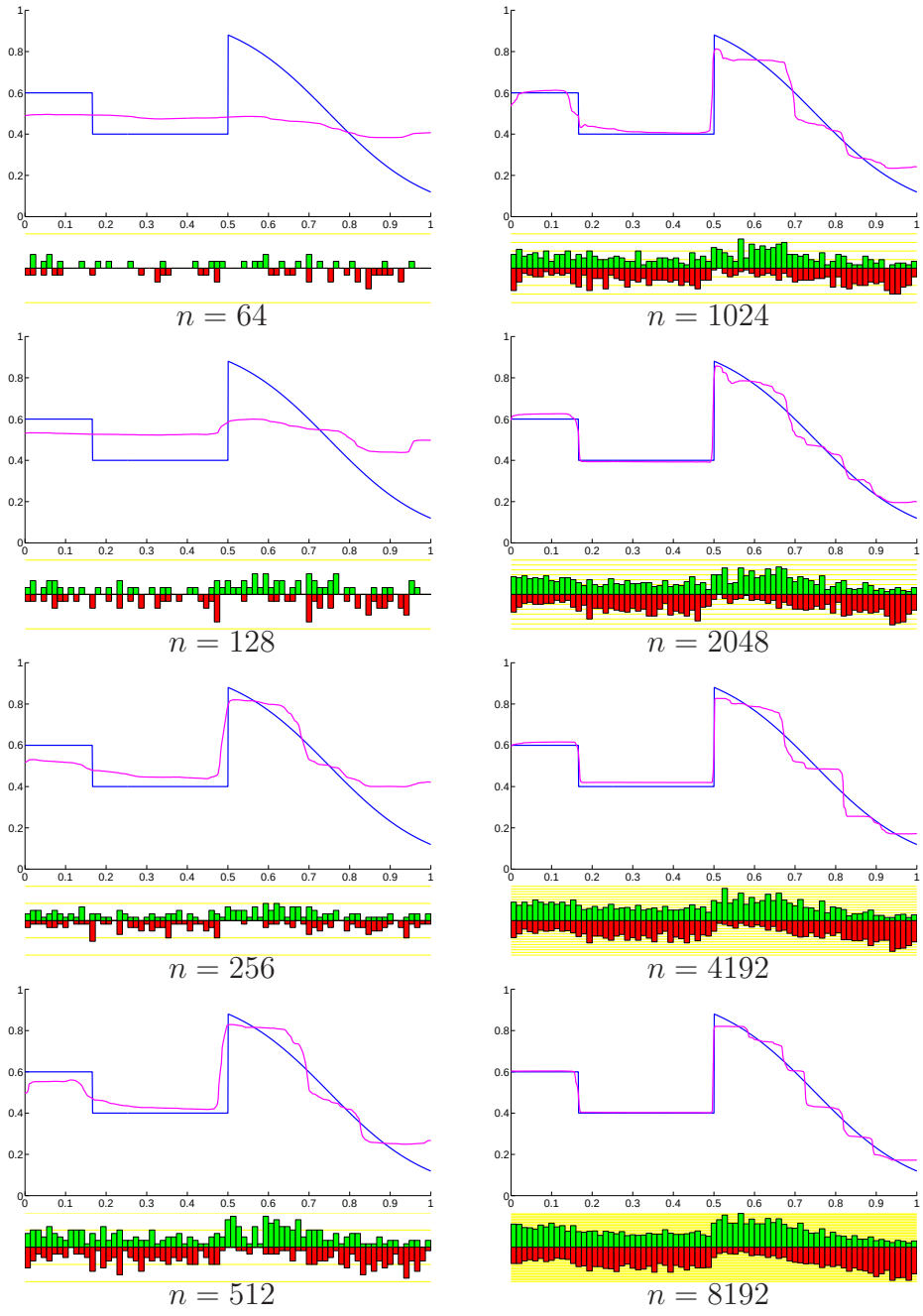

Figure 4.8.a: *The Effect of Sample Size:* The posterior mean is computed as sample size n runs through a wide range

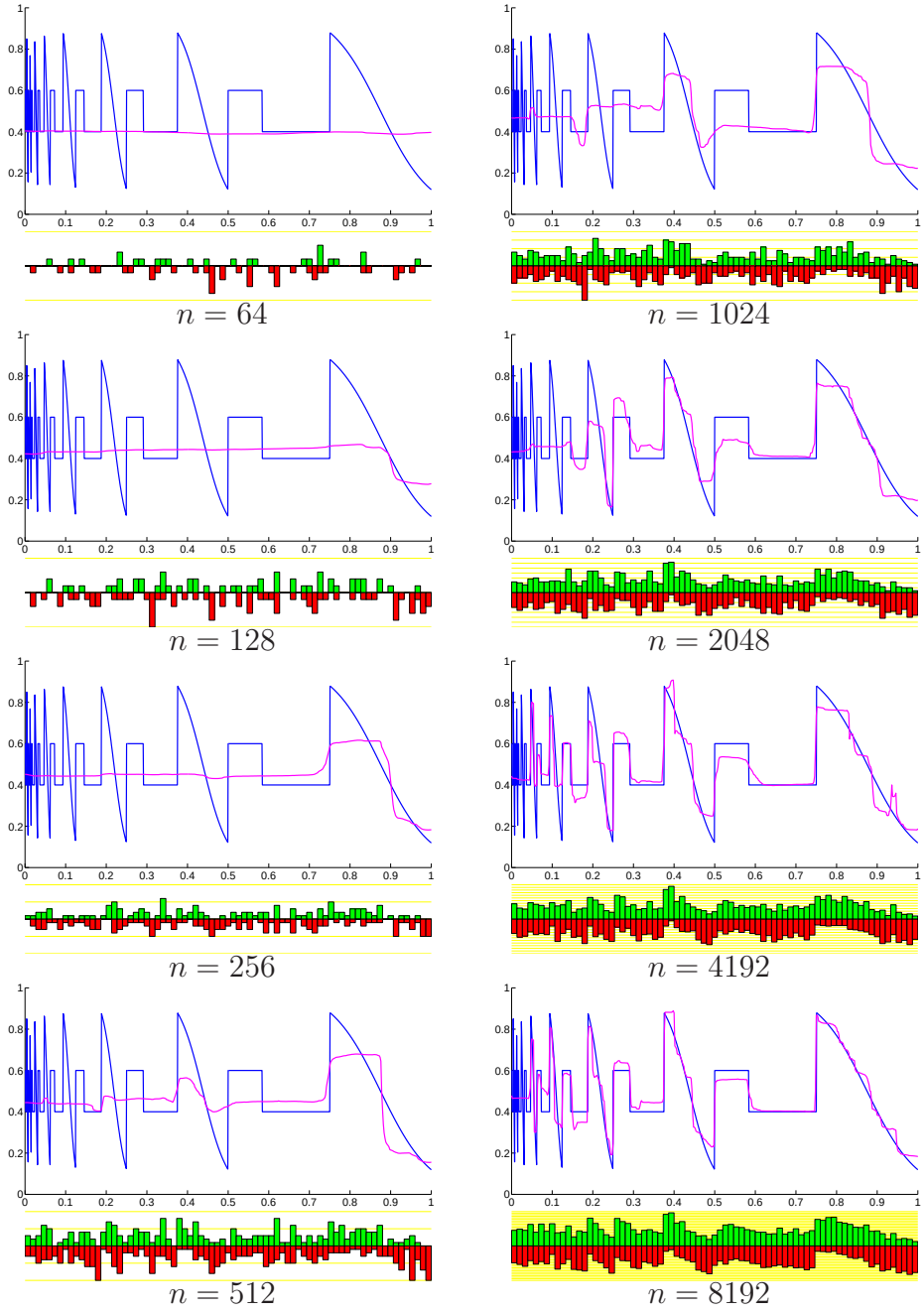

Figure 4.9.a: *The Effect of Sample Size: a Harder Example* The posterior mean is computed as sample size n runs through a wide range on this more challenging problem

left. Importantly, for the larger n, the posterior mean concentrates on models with many more splits than it did for the easier data since. Presumably this is because more complex models are necessary to increase the likelihood. To some extent this makes sense because even though this example is more complex than the previous one, some aspects of the regression function are relatively easy to detect (e.g. the large jump) and there are more of these features available for analysis. It is interesting to compare the results with the former dataset for $n = 4096$ with the results *on the right half* of the current dataset for $n = 8192$. They should be quite comparable because the regression function for this more complicated model on $[\frac{1}{2}, 1]$ is just a rescaled version of the original model, and in both cases there should be about 4096 datapoints available. To appearances, the two results are quite similar except that the latter result does not do as well on the small interval from 0.5 to around 0.6. Perhaps it is being confused by the low success probability region immediately to the left of 0.5. Conducting a similar comparison between the original $n = 1024$ example the right half of the $n = 2048$ example, the latter result seems substantially inferior. On the other hand, it is not so bad when compared with the original result on experimental run 5. Furthermore, the performance of the posterior mean is quite good considering the overall increase in the difficulty of this problem. Even more amazing, considering the popular state-of-the-art methods, it does all of this without any tuning or cross-validation.

Chapter 5

Consistency

This chapter establishes conditions under which the prior for one-dimensional classification that was introduced in section 1.2 is a consistent estimator of the true regression function. The consistency of the posterior is proven using a result by Barron, Schervish, and Wasserman [2]. Before reviewing their theorem, I pause to introduce some notation. I also present a lemma that shows that the original conditions given in their theorem are equivalent to some others that may be easier to check. Finally, I specify the prior π more formally and complete a proof of consistency.

5.1 Notation and the Basic Theorem

Let $\mathcal{F}$ be the class of all (Borel) measurable functions $f : [0,1] \to \{0,1\}$. Write μ for the uniform distribution on $[0,1]$ and η for counting measure on the set $\{0,1\}$. Write $\mathcal{Z}$ for the product space $[0,1] \times \{0,1\}$, and call the product measure ν. To any $f \in \mathcal{F}$ there is a corresponding density on $\mathcal{Z}$ with respect to ν that we denote by $\widetilde{f}$:

$$\widetilde{f}(x,y) = f(x)\mathbf{1}_{y\,=\,1} + (1 - f(x))\mathbf{1}_{y\,=\,0} \tag{5.1}$$

For notational convenience, we may write either $\widetilde{f}(x,y)$ for $x \in [0,1]$ and $y \in \{0,1\}$ or $\widetilde{f}(z)$ for $z = (x,y) \in \mathcal{Z}$. Write $F = F_f$ for the distribution on $\mathcal{Z}$ whose density with respect to ν is $\widetilde{f}$. In words, the consequence of this construction is that sampling a point Z from F is the same as choosing an X uniformly on $[0,1]$ and then (conditionally on $X = x$) determining Y by flipping an "$f(x)$" coin.

Write $\widetilde{\mathcal{F}}$ for the class of densities formed by considering $\widetilde{f}$ for every $f \in \mathcal{F}$.

Let d denote the Hellinger distance on $\widetilde{\mathcal{F}}$:

$$d(\widetilde{f}, \widetilde{f}') = \Big\{ \int \big(\widetilde{f}(z)^{\frac{1}{2}} - \widetilde{f}'(z)^{\frac{1}{2}} \big) \nu(dz) \Big\}^{\frac{1}{2}} \tag{5.2}$$

And let D denote the Kullback-Leibler discrepancy on $\widetilde{\mathcal{F}}$ (employing the usual convention that the integrand is interpreted as 0 whenever $\widetilde{f}(z) = 0$):

$$D(\widetilde{f}, \widetilde{f}') = \int \log \frac{\widetilde{f}(z)}{\widetilde{f}'(z)} \widetilde{f}(z) \nu(dz) \tag{5.3}$$

We are concerned with posterior consistency and so the mass that the prior or posterior ascribes to certain small sets containing the true parameter $\widetilde{f}_0$ is of interest. For any $\epsilon > 0$, we define two such "neighborhoods:"

$$K_\epsilon = \{ \widetilde{f} \in \widetilde{\mathcal{F}} \mid D(\widetilde{f}_0, \widetilde{f}) \leq \epsilon \} \tag{5.4}$$

$$H_\epsilon = \{ \widetilde{f} \in \widetilde{\mathcal{F}} \mid d(\widetilde{f}_0, \widetilde{f}) \leq \epsilon \} \tag{5.5}$$

The "richness" of the parameter space is also an important quantity; to make this precise we supply the following definitions.

Definition 1. Consider a class of functions $\mathcal{A}$ that are densities with respect to dominating measure ν. We say that the collection of functions $\{\widetilde{f}_1^U, \ldots, \widetilde{f}_r^U\}$ is a δ-upper bracketing of $\mathcal{A}$ if:

1. for every $a \in \mathcal{A}$ there exists i such that $a \leq \widetilde{f}_i^U$ a.e. $[\nu]$

2. every $\widetilde{f}_i^U$ satisfies $\int \widetilde{f}_i^U(z)\nu(dz) \leq 1+\delta$

Furthermore, write $\mathcal{H}(\mathcal{A}, \delta)$ for the δ-upper metric entropy of $\mathcal{A}$ which is the logarithm of the size of the smallest possible δ-upper bracketing of $\mathcal{A}$ (infinity, if no finite bracketings exist).

The following result can now be stated. It gives conditions on the prior that are sufficient to ensure that the posterior will concentrate on H_ϵ for any $\epsilon > 0$.

Theorem 4 (Barron, Schervish, and Wasserman 1999). *Let ν be a σ-finite measure on a measurable space $(\mathcal{Z}, \mathcal{B})$, where the σ-field $\mathcal{B}$ is separable. Let π be a probability distribution on $\widetilde{\mathcal{F}}$, a class of probability densities with respect to ν. Endow $\widetilde{\mathcal{F}}$ with the Borel σ-field induced by the Hellinger metric d. Let $\widetilde{f}_0$ be a certain chosen density with respect to ν and write F_0 for the corresponding distribution on $\mathcal{Z}$. Let $Z_1, Z_2, \ldots$ be drawn iid from F_0. In the notations explained above, further assume that for every $\epsilon > 0$:*

1. *$\pi(K_\epsilon) > 0$*

2. *There exists a sequence $\{\mathcal{A}_n\}_{n=1}^\infty$ of measurable subsets of $\widetilde{\mathcal{F}}$ and positive, real numbers d_1, d_2, c, and δ such that:*

   (a) *$\pi(\mathcal{A}_n^c) \leq d_1 \exp(-d_2 n)$ for all n sufficiently large*

   (b) *$\mathcal{H}(\mathcal{A}_n, \delta) \leq cn$ for all n sufficiently large*

   (c) *$c < \left((\epsilon - \sqrt{\delta})^2 - \delta\right)/2$*

   (d) *$\delta < \epsilon^2/4$*

Then, with probability 1 [under F_0^∞ measure], Bayes theorem applies for all n; i.e. for any measurable $B \subset \widetilde{\mathcal{F}}$ and any n:

$$\pi(B|z_1, \ldots, z_n) = \frac{\int_B \prod_{i=1}^n \widetilde{f}(z_i)\pi(d\widetilde{f})}{\int_{\widetilde{\mathcal{F}}} \prod_{i=1}^n \widetilde{f}(z_i)\pi(d\widetilde{f})}$$

And for any $\epsilon > 0$*:*

$$\lim_{n\to\infty} \pi(H_\epsilon | z_1, \ldots, z_n) = 1$$

The second condition of this theorem seems rather technical. To restate this theorem in simpler terms we prove an elementary lemma which shows that these conditions (item *I* in the lemma) can be expressed in three other equivalent forms.

Lemma 5. *The following conditions, given in the notation defined above, are equivalent:*

I For all $\epsilon > 0$*, there exists* $\{\mathcal{A}_n\}_{n=1}^{\infty}$*, a sequence of measurable subsets of* $\widetilde{\mathcal{F}}$ *and positive, real numbers* d_1, d_2, c*, and* δ *such that:*

- *a* $\pi(\mathcal{A}_n^c) \leq d_1 \exp(-d_2 n)$ *for all* n *sufficiently large*
- *b* $\mathcal{H}(\mathcal{A}_n, \delta) \leq cn$ *for all* n *sufficiently large*
- *c* $c < \left((\epsilon - \sqrt{\delta})^2 - \delta\right)/2$
- *d* $\delta < \epsilon^2/4$

II There exists a sequence of positive real numbers $\{\delta^i\}_{i=1}^{\infty}$*, with* $\delta^i \downarrow 0$*, and* $\{\{\mathcal{A}_n^i\}_{n=1}^{\infty}\}_{i=1}^{\infty}$*, a sequence of sequences of measurable subsets of* $\widetilde{\mathcal{F}}$ *such that:*

- *a for all* i*,* $\limsup_n \ \log\left(\pi\left((\mathcal{A}_n^i)^c\right)\right)/n < 0$
- *b* $\limsup_i \limsup_n \mathcal{H}(\mathcal{A}_n^i, \delta^i)/n = 0$

III For all $\epsilon > 0$ *there exists* $\{\mathcal{A}_n\}_{n=1}^{\infty}$*, a sequence of measurable subsets of* $\widetilde{\mathcal{F}}$ *and a positive, real number* $\delta \leq \epsilon$ *such that:*

- *a* $\limsup_n \ \log\left(\pi\left((\mathcal{A}_n)^c\right)\right)/n < 0$
- *b* $\limsup_n \mathcal{H}(\mathcal{A}_n, \delta)/n \leq \epsilon$

IV For all $\epsilon > 0$ there exists $\{\mathcal{A}_n\}_{n=1}^{\infty}$, a sequence of measurable subsets of $\widetilde{\mathcal{F}}$ such that:

a $\limsup_n \ \log(\pi((\mathcal{A}_n)^c))/n < 0$

b $\limsup_n \mathcal{H}(\mathcal{A}_n, \epsilon)/n \leq \epsilon$

Proof. I $\implies$ II:

First, notice that the condition from *I.a* that "there exist $d_1 > 0$ and $d_2 > 0$ so that $\pi(A_n^c) \leq d_1 \exp(-d_2 n)$ for n sufficiently large" implies that $\limsup_n \log(\pi(\mathcal{A}_n^c))/n \leq -d_2 < 0$. Conversely, if $\limsup_n \log(\pi(\mathcal{A}_n^c))/n = \alpha < 0$, then for all sufficiently large n, $\log(\pi(\mathcal{A}_n^c))/n \leq \alpha/2$ and $\pi(\mathcal{A}_n^c) \leq \exp(\frac{\alpha}{2}n)$ for all sufficiently large n.

Choose a sequence of $\epsilon^i > 0$, $\epsilon^i \downarrow 0$ and use I to establish the existence of $\{\mathcal{A}_n^i\}, d_1^i, d_2^i, c^i, \delta^i$ satisfying I for ϵ^i. Since $\delta^i < (\epsilon^i)^2/4 \downarrow 0$, we can find a subsequence on which $\delta^i \downarrow 0$. Without loss of generality, assume we already have such a subsequence. The above reasoning, then, establishes *II.a* for this sequence. Since *I.a* implies that $\limsup_n \mathcal{H}(\mathcal{A}_n^i, \delta^i)/n \leq c^i$ for all i, to establish *II.b*, we need only show that $\limsup_i c^i = 0$. Condition *I.c* constrains $c^i < \left((\epsilon^i - \sqrt{\delta^i})^2 - \delta^i\right)/2$. Viewing this as a function of δ^i, notice that it is monotonically decreasing on $[0, (\epsilon^i)^2/4]$ so that, necessarily, $c^i < (\epsilon^i)^2/2$, the value of this function at $\delta^i = 0$.

II $\implies$ I:

Consider some $\epsilon > 0$. Note that $\left((\epsilon - \sqrt{\delta^i})^2 - \delta^i\right)/2 \uparrow \epsilon^2/2$ as $i \to \infty$. Find an i sufficiently large so that this expression is at least $\epsilon^2/3$. Then find a subsequent i sufficiently large so that $\limsup_n \mathcal{H}(\mathcal{A}_n^i, \delta^i)/n \leq \epsilon^2/5$. Choose $c = \epsilon^2/4$, and $\delta = \delta^i$, and I is proven.

II $\iff$ III: This is a straightforward exercise in nitpicking.

III $\iff$ IV: Observe from the definition that $\mathcal{H}(\mathcal{A}_n, \delta)$ is a non-increasing function of δ.

□

Theorem 6 (Corollary to Barron, Schervish, and Wasserman 1999). *Let ν be a σ-finite measure on a measurable space $(\mathcal{Z}, \mathcal{B})$, where the σ-field $\mathcal{B}$ is separable. Let π be a probability distribution on $\widetilde{\mathcal{F}}$, a class of probability densities with respect to ν. Endow $\widetilde{\mathcal{F}}$ with the Borel σ-field induced by the Hellinger metric d. Let $\widetilde{f}_0$ be a certain chosen density with respect to ν and write F_0 for the corresponding distribution on $\mathcal{Z}$. Let $Z_1, Z_2, \ldots$ be drawn iid from F_0. In the notations explained above, further assume that for every $\epsilon > 0$:*

1 $\pi(K_\epsilon) > 0$

There exists a sequence $\{\mathcal{A}_n\}_{n=1}^{\infty}$ of measurable subsets of $\widetilde{\mathcal{F}}$, so that:

2 $\limsup_n \ \log(\pi((\mathcal{A}_n)^c))/n < 0$

3 $\limsup_n \ \mathcal{H}(\mathcal{A}_n, \epsilon)/n \leq \epsilon$

Then, with probability 1 [under F_0^∞ measure], Bayes theorem applies for all n; i.e. for any measurable $B \subset \widetilde{\mathcal{F}}$ and any n:

$$\pi(B|z_1, \ldots, z_n) = \frac{\int_B \prod_{i=1}^n \widetilde{f}(z_i)\pi(d\widetilde{f})}{\int_{\widetilde{\mathcal{F}}} \prod_{i=1}^n \widetilde{f}(z_i)\pi(d\widetilde{f})}$$

And for any $\epsilon > 0$:

$$\lim_{n\to\infty} \pi(H_\epsilon|z_1, \ldots, z_n) = 1$$

In words, this is what we have required: 1) the prior must put positive mass on all Kullback-Leibler neighborhoods of $\widetilde{f}_0$. 2) We must be able to choose an increasing sequence of subsets of the parameter space so that the n'th of these sets captures all but exponentially much of the prior mass 3) This sequence must not grow in "complexity" too quickly. We conclude that the posterior will concentrate on the subset of the parameter space which is Hellinger close to the true parameter.

5.2 Specification of the Prior

We describe π as a distribution on $\widetilde{\mathcal{F}}$ by means of first describing a parametric prior π'. Let κ be a distribution on the positive integers. Let $\Theta_k = \{(k, \mathbf{v}, \mathbf{s}) \ : \ \mathbf{v} \in [0,1]^{k-1}, \mathbf{s} \in [0,1]^k\}$ and let $\Theta = \cup_{k=1}^{\infty}\Theta_k$. Let π' be the distribution on Θ, that can be described by first picking K according to κ; and then, conditional on $K = k$, picking a point $\theta \in \Theta$ uniformly from Θ_k. That is, $\mathbf{v}$ and $\mathbf{s}$ are chosen independently and uniformly from the appropriate unit cubes. Let $\mathbf{v}_{(i)}$ denote the i'th ordered value of $\mathbf{v}$. Now to any $\theta \in \Theta$ associate the function $f_\theta \in \mathcal{F}$ given by the following construction. For $k > 1$, let $I_1 = [0, \mathbf{v}_{(1)}), I_2 = [\mathbf{v}_{(2)}, \mathbf{v}_{(3)}), \ldots, I_{k-1} = [\mathbf{v}_{(k-2)}, \mathbf{v}_{(k-1)}), I_k = [\mathbf{v}_{(k-1)}, 1]$. If $k = 1$, just let $I_1 = [0,1]$. Take $f_\theta(x) = \sum_{i=1}^{k} \mathbf{s}_i \mathbf{1}_{x \in I_i}$. Finally, then, to draw $\widetilde{f}$ from π, draw θ from π', construct f_θ, and form $\widetilde{f}_\theta$ as in Equation 5.1.

5.3 A Consistency Proof

This section proves that the prior π just described is consistent in the sense that, under repeated sampling, the posterior mass will concentrate on a Hellinger neighborhood H_ϵ for any $\epsilon > 0$. The proof uses three lemmas to establish the conditions of Theorem 6.

The first lemma shows that π puts mass on all Kullback-Leibler neighborhoods K_ϵ.

Lemma 7. *Let π be the prior distribution on $\widetilde{\mathcal{F}}$, the class of densities w.r.t ν, that was described above. Assume that κ, the hierarchy prior, assigns positive mass to every natural number. Let $\widetilde{f}_0$ be an arbitrary density in $\widetilde{\mathcal{F}}$. Then, for any $\epsilon > 0$, $\pi(K_\epsilon) > 0$.*

Proof. Let f_0 be the corresponding function in $\mathcal{F}$. For some $0 < \delta_1 < \frac{1}{16}$, to be determined later, let:

$$\mathcal{G} = \{g \in \mathcal{F} : \ \ ||g - f_0||_1 \leq 2\delta_1 \text{ and } \delta_1 \leq g(x) \leq 1 - \delta_1 \ \ (\forall x)\}$$

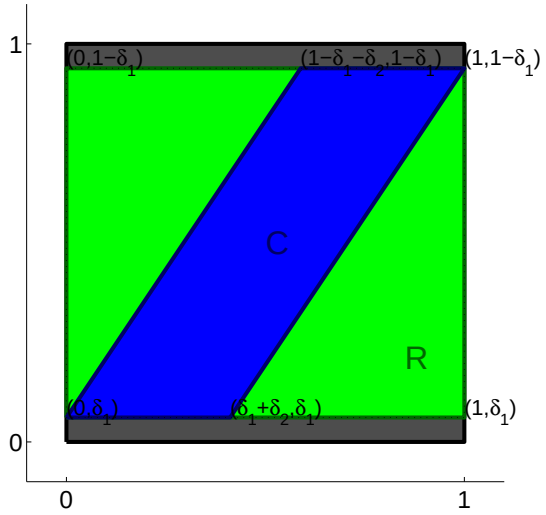

Figure 5.3.a: The sets C and R

Let $A_g = \{x \in [0,1] \ : \ |g(x) - f_0(x)| \leq \delta_2\}$ where $\delta_2 := \sqrt{2\delta_1}$. By Chebyshev's inequality, for any $g \in \mathcal{G}$, the Lebesgue measure of ${A_g}^c$ is smaller than δ_2.

In other words, as shown in Figure 5.3.a, if $x \in A_g$, this restricts the pair $(f_0(x), g(x))$ to lie in the convex set C (blue) whose extreme points are $(0, \delta_1)$, $(\delta_1+\delta_2, \delta_1)$, $(1, 1-\delta_1)$, $(1-\delta_1-\delta_2, 1-\delta_1)$. For $x \notin A_g$, we still have a restriction on $g(x)$, so that the pair $(f_0(x), g(x))$ lies in the rectangle R (green) with vertices $(0, \delta_1)$, $(1, \delta_1)$, $(1, 1-\delta_1)$, $(0, 1-\delta_1)$. Let $D_1(p, p')$ denote the Kullback-Leibler discrepancy between the $Bernoulli(p)$ and $Bernoulli(p')$ distributions.

$$D_1(p, p') := p \log \frac{p}{p'} + (1-p) \log \frac{1-p}{1-p'}$$

Then for any $g \in \mathcal{G}$ we can bound $D(\widetilde{f}_0, \widetilde{g})$, as follows:

$$D(\widetilde{f}_0, \widetilde{g}) = \int_{A_g} D_1(f_0(x), g(x)) + \int_{{A_g}^c} D_1(f_0(x), g(x)) \tag{5.6}$$

$$\leq \int_{A_g} \sup_{(a,b)\in C} D_1(a,b) + \int_{{A_g}^c} \sup_{(a,b)\in R} D_1(a,b) \tag{5.7}$$

and because $D_1(a,b)$ is convex in the pair (a,b) [13], the supremum is achieved at the vertices so that the above is bounded by:

$$\leq \max_{(a,b)\in\{\text{vertices of } C\}} D_1(a,b) + \delta_2 \max_{(a,b)\in\{\text{vertices of } R\}} D_1(a,b) \tag{5.8}$$

Using the symmetry $D_1(a,b) = D_1(1-a, 1-b)$

$$\leq \max\left(D_1(0,\delta_1), D_1(\delta_1+\delta_2,\delta_1)\right) + \delta_2 \max\left(D_1(0,\delta_1), D_1(0,1-\delta_1)\right) \tag{5.9}$$

$$= \max\Bigg(-(1-\delta_1)\log(1-\delta_1), \\ (\delta_1+\delta_2)\log(\frac{\delta_1+\delta_2}{\delta_2}) + (1-\delta_1-\delta_2)\log(\frac{1-\delta_1-\delta_2}{1-\delta_1})\Bigg) \\ + \delta_2(-\log(\delta_1)) \tag{5.10}$$

$$\leq \max\left(-\log(1-\delta_1), 2\log(2)\delta_2\right) + \delta_2(-\log(\delta_1)) \tag{5.11}$$

For small δ_1, the last term is the most important. To simplify the first term, verify that for $0 \leq \delta \leq \frac{1}{2}$, $-\log(1-\delta) \leq \delta + \delta^2 \leq \frac{3}{2}\delta$

$$\leq \max\left(\frac{3}{2}\delta_1, 2\log(2)\delta_2\right) + \delta_2(-\log(\delta_1)) \tag{5.12}$$

$$\leq \frac{3}{2}\delta_2 + \delta_2(-\log(\delta_1)) \tag{5.13}$$

$$= \sqrt{2\delta_1}(-\log(\delta_1) + \frac{3}{2}) \tag{5.14}$$

Which tends to zero as $\delta_1 \downarrow 0$. For a sufficiently small choice of δ_1, then:

$$\widetilde{\mathcal{G}} = \{\widetilde{g} \ : \ g \in \mathcal{G}\} \subset K_\epsilon = \{\widetilde{f} \in \widetilde{\mathcal{F}} \ : \ D(\widetilde{f}_0, \widetilde{f}) < \epsilon\}.$$

It remains to show that $\widetilde{\mathcal{G}}$ has positive prior mass. To do this, we first find a step function $g_0 \in \mathcal{G}$ which approximates f_0 and then show that $\mathcal{G}^*$, a set of perturbations of g_0, remains in $\mathcal{G}$ and that $\widetilde{\mathcal{G}^*}$ has positive prior mass.

Observe that since f_0 is Lebesgue-measurable there exist two increasing sequences of step functions h_i^+ and h_i^- for which $||f_0 - (h_i^+ - h_i^-)||_1 \to 0$. This is, in fact, the basis of a common construction of the Lebesgue integral [42]. Consequently, we can find a step function $h \in \mathcal{F}$ for which $||f_0 - h||_1 \leq \delta_1/2$. Let g_0 be the function obtained by modifying h so that it always remains in $[\delta_1, 1-\delta_1]$, i.e. $g_0(x) = \max(\min(h(x), 1-\delta_1), \delta_1)$, so that $g_0 \in \mathcal{G}$ and $||f_0 - g_0||_1 \leq \frac{3}{2}\delta_1$.

Now, parameterize g_0 in the manner of section 5.2 (changing it at a set of measure 0 if necessary) so that $k < \infty$ is the number of locally constant regions in g_0 and the vectors $\mathbf{v} = (v_i)_{i=1}^{k-1}$ and $\mathbf{s} = (s_i)_{i=1}^{k}$ signify the change-points and success probabilities of g_0.

If we then perturb each s_i towards the value $\frac{1}{2}$ by no more than $\delta_1/4$ we obtain a new function g' which remains in $\mathcal{G}$ and satisfies $||g_0 - g'||_1 \leq \delta_1/4$. If, in addition, we perturb the v_i's by no more than $\delta_1/(4(k-1))$ we obtain g'' which also remains in $\mathcal{G}$, satisfying $||g_0 - g''||_1 \leq \delta_1/2$ or $||f_0 - g''||_1 \leq 2\delta_1$.

Denote the class of functions thus obtained by $\mathcal{G}^*$. Then $\pi(\widetilde{\mathcal{G}}) \geq \pi(\widetilde{\mathcal{G}^*}) \geq \kappa\{k\}(\delta_1/4)^k(\delta_1/(2(k-1)))^{k-1} > 0$. □

Remark 1. A shorter proof can be based on the martingale sequence formed from conditional expectations of f_0.

For the second lemma, consider $\mathcal{F}_m$, the class of functions which are constant except possibly for as many as $m-1$ change-points.

Definition 2. Using the notation from section 5.2, let $\mathcal{F}_m = \{f_\theta \ : \ \theta \in \Theta_m\}$. Let $\widetilde{\mathcal{F}}_m$ be the associated class of densities with respect to ν:

$$\widetilde{\mathcal{F}}_m = \{\widetilde{f}(x,y) = f(x)\mathbf{1}_{y\,=\,1} + (1-f(x))\mathbf{1}_{y\,=\,0} : f \in \mathcal{F}_m\}$$

Lemma 8. *The δ-upper bracketing entropy of the class $\widetilde{\mathcal{F}}_m$ is no more than $(2m-1)\log(\lceil m/\delta \rceil)$.*

Proof. Fix a, b positive integers. Partition $[0,1]$ into a equal intervals $I_1, \ldots, I_a$. Consider the class $H_{m;a,b}$ of all functions $\widetilde{f}^U$ that can be formed in the following way: Choose some $m-1$ of these intervals. Let C be the union of the chosen intervals and let $C_1, \ldots, C_k$ be the nonempty subintervals of $[0,1]$ formed by subtracting C ($k \leq m$). Finally choose $B_1, \ldots B_k \in \{1, \ldots, b\}$. Construct the function $\widetilde{f}^U : [0,1] \times \{0,1\} \mapsto [0,1]$ by:

$$\widetilde{f}^U(x,y) = \begin{cases} 1 & \text{if } x \in C \\ B_i/b & \text{if } x \in C_i \text{ and } y = 1 \\ 1-(B_i-1)/b & \text{if } x \in C_i \text{ and } y = 0 \end{cases}$$

It is easy to see that, by appropriate choices, for any $\widetilde{f} \in \widetilde{\mathcal{F}}_m$ we can find an $\widetilde{f}^U \in H_{m;a,b}$ that is greater than or equal to it globally. Furthermore, the integral of $\widetilde{f}^U$ is less than or equal to $1 + 1/b + (m-1)/a$. The size of $H_{m;a,b}$ is $\leq \binom{a}{m-1} b^m \leq a^{m-1}b^m$. By choosing $a = b = \lceil m\delta^{-1} \rceil$, we have shown that $\exp(\mathcal{H}(\widetilde{\mathcal{F}}_m, \delta)) \leq (\lceil m\delta^{-1} \rceil)^{2m-1}$. □

The final lemma establishes a result about the tail of *Poisson* random variables.

Lemma 9. *If* $K \sim$ *Poisson*(λ) *for some* $\lambda > 0$*, then for any* $k > \lambda$*:*

$$\mathbf{P}(K \geq k) \leq e^{-\lambda} \frac{\lambda^k}{k!} \left(\frac{k}{k - \lambda} \right)$$

And, consequently, for any $0 < \beta < 1$ *there is a* k_0 *sufficiently large so that for every* $k \geq k_0$*,* $\mathbf{P}(K \geq k) \leq \exp(-\beta k \log(k))$*.*

Proof.

$$\begin{aligned} \mathbf{P}(K \geq k) &= e^{-\lambda} \sum_{i=k}^{\infty} \frac{\lambda^i}{i!} \\ &= e^{-\lambda} \frac{\lambda^k}{k!} \sum_{i=k}^{\infty} \left(\frac{\lambda}{k} \right)^{i-k} \left[\frac{k! k^{i-k}}{i!} \right] \end{aligned}$$

Now, the bracketed term is no more than 1 and we are left with a geometric series whose factor, λ/k is less than 1 by our assumption so that:

$$\begin{aligned} \mathbf{P}(K \geq k) &\leq e^{-\lambda} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{k} \right)^{-1} \\ &= e^{-\lambda} \frac{\lambda^k}{k!} \left(\frac{k}{k - \lambda} \right) \end{aligned}$$

So that for k sufficiently large, from Stirling's formula:

$$\begin{aligned} \log(\mathbf{P}(K \geq k)) &\leq -\log(k!) - \lambda + k \log(\lambda) + \log \left(\frac{k}{k - \lambda} \right) \\ &\leq -\beta k \log(k) \end{aligned}$$

□

The main result can now be established.

Theorem 10. *Suppose that the prior π, described in section 5.2 is based on the hierarchy prior κ. Suppose that κ gives positive probability to every natural number and that its tail satisfies: $\kappa(\{k \in \mathbb{N} \ : \ k \geq j\}) \leq \exp(-\beta j \log(j))$ for all j sufficiently large and some $\beta > 0$. Let f_0 be an arbitrary measurable function from $[0,1]$ into $[0,1]$. Suppose that $Z_1, Z_2, \ldots$ are drawn iid from the distribution F_0, which has density $\widetilde{f}_0$ with respect to ν. Equivalently, suppose that the Z_i's ($Z_i = (X_i, Y_i)$) are drawn as follows: the X_i's are drawn independently and uniformly from $[0,1]$; and, conditional on $X_i = x_i$, Y_i is an independent Bernoulli$(f_0(x_i))$ random variable. Then, in the notation of section 5.1, for any $\epsilon > 0$, $\pi(H_\epsilon | z_1, \ldots, z_n) \to 1$ as $n \to \infty$ [${F_0}^\infty$-a.s.]. Specifically, for any $1 \leq p < \infty$, and any $\epsilon > 0$, the posterior mass on the set $\{\widetilde{f} \in \widetilde{\mathcal{F}} \ : \ ||f - f_0||_p < \epsilon\}$ tends to 1 as $n \to \infty$ [${F_0}^\infty$-a.s.].*

Proof. Let $m(n) := \lfloor \alpha n / \log(n) \rfloor$. Choose the sequence $\{\mathcal{A}_n\}$ as $\mathcal{A}_n = \widetilde{\mathcal{F}}_{m(n)}$. Then, by Lemma 8, $\mathcal{H}(\mathcal{A}_n, \epsilon) \leq (2m(n) - 1)\log(\lceil m(n)/\epsilon \rceil)$. Observe that $m(n) \uparrow \infty$ as $n \to \infty$. For n large enough, then:

$$\begin{aligned}
\mathcal{H}(\mathcal{A}_n, \epsilon) &\leq 2m(n)\log(\alpha n/\log(n)\epsilon^{-1} + 1) \\
&\leq 2\alpha n/\log(n)[\log(n/\log(n)) + \log(\alpha) - \log(\epsilon) + 1] \\
&\leq 3\alpha n \frac{\log(n/\log(n))}{\log(n)}
\end{aligned}$$

Choosing $\alpha = \epsilon/3$, we have proven that $\limsup \mathcal{H}(\mathcal{A}_n, \epsilon)/n \leq \epsilon$.

Now calculate that for all sufficiently large n,

$$\begin{aligned}
-\log(\pi({\mathcal{A}_n}^c)) &= -\log(\kappa(\{k \in \mathbb{N} \ : \ k \geq m(n) + 1\})) \\
&\geq \beta[m(n) + 1]\log(m(n) + 1) \\
&\geq \beta[\alpha n/\log(n)]\log(\alpha n/\log(n)) \\
&= \alpha\beta n \frac{\log(n) - \log(\log(n)) + \log(\alpha)}{\log(n)} \\
&\geq \frac{1}{2}\alpha\beta n
\end{aligned}$$

Consequently, $\limsup \log(\pi(\mathcal{A}_n{}^c))/n \leq -\frac{1}{2}\alpha\beta < 0$.

Finally, recall that Lemma 7 shows that for any $\epsilon > 0$, $\pi(K_\epsilon) > 0$. Apply Theorem 6 to complete the proof.

The statement about $||\cdot||_p$ follows from the equivalence of the $\mathcal{L}^p$ and Hellinger metrics for bounded densities. This fact and other useful inequalities about Hellinger distance d (aka Jeffrey's distance) are reviewed in [22, section 5.8 and excercise 5.7] which states:

$$d(f,g)^2 \leq ||f-g||_1 \leq 2d(f,g) \tag{5.15}$$

Finally, the equivalence of the $\mathcal{L}^1$-norm and the $\mathcal{L}^p$ norm for $1 \leq p < \infty$ and for all densities uniformly bounded by a constant, is well known. □

Remark 2. Applying Lemma 9 shows that the preceding theorem applies to any hierarchy prior κ whose tail behaves like that of a *Poisson*(λ), for any $\lambda > 0$. The theorem does not apply to the case in which κ is *Geometric*.

Remark 3. The restrictions on the tail of the prior occur because condition 3 in Theorem 6 requires that the "sieve" sets $\mathcal{A}_n$ do not grow in "size" too quickly as n grows. The choice $\mathcal{A} = \widetilde{\mathcal{F}}_{m(n)}$ made in the proof of this theorem is (essentially) the fastest rate of growth that this situation permits (in the absence of better bracketing estimates). Accordingly, condition 2 of Theorem 6 requires that the tail of the prior drops off somewhat faster than a *Geometric* distribution.

Remark 4. For further discussion of how to interpret this result, please see the discussion in chapter 6.

Chapter 6

Discussion of Consistency Results

It is challenging to suggest what the practical consequences (if any) of Theorem 10 are. It proves that under the modeling set up with *iid* observations that has been considered throughout this thesis, that the posterior of the random split prior π is a consistent estimate of any measurable true regression function *if* the tail of the prior decays at least as fast as $\exp(-\beta j \log(j))$ for some $\beta > 0$. Perhaps it would be wise not to over interpret the result. After all, it only supplies a sufficient condition for consistency and does not either establish that *Geometric* priors lead to inconsistent estimators or that *Poisson*-like priors are a good (i.e. practical) idea. Additionally, it attempts to prove consistency in a fairly strong sense: that the posterior mass concentrates on Hellinger neighborhoods of the truth. Weaker consistency results, say that the posterior mean be $\mathcal{L}^2$ consistent, might go through under milder assumptions.

6.1 Consideration of the Diaconis and Freedman Results

Judging from the results of Diaconis and Freedman for their prior, as discussed in subsection 2.2.1, it might, in fact, be the case that the posterior is consistent for any hierarchy prior κ (with full support), so long as f_0 is not the constant function $\frac{1}{2}$. That is, it might be that the only situation in which π is inconsistent (even using a "poor" choice of κ) is when there is no *real* pattern at all in the data because every coin flip was fair. Note that if this were our only concern, namely that the estimates might be inconsistent under some specific finite collection of possible scenarios, this could be easily addressed (albeit in a decidedly non-Bayesian way), by choosing a prior that puts point mass on the troublesome cases. Consistency for these exceptions would then be guaranteed by a suitable application of Doob's result [29]. Interestingly, it could destroy consistency for other cases. An example of such a mixture is given by Diaconis and Freedman [31]. Of course, a true subjective Bayesian would never change his or her prior in this way. Rather, unless these cases actually are a subjective impossibility, a purist would merely see this as an explication of why their *true* prior (that they are perhaps still in the process of articulating) differs from the former one.

6.2 An Experiment to Check a Worrisome Case

It seems reasonable to try and test for the possibility of inconsistency when the true function $f_0 \equiv \frac{1}{2}$ by running a simulation experiment. To do this, generate an increasing sequence of data sets over a range of sizes that are drawn from the $f_0 \equiv \frac{1}{2}$ distribution. That is, since there is no true signal at all in the data, it will be interesting to see how the posterior mean responds. For a *Poisson* prior, the posterior mean should (at least eventually) settle down, but will this hold for a *Geometric* prior? Indeed, the results in figures 6.2.a, and 6.2.b

indicate that both estimates correctly identify the null case, whether using a $Poisson(5)$ prior or a $Geometric(\frac{1}{2})$ one respectively. A new feature of these figures is the small histogram included on each one. It is a histogram of the posterior on the number of steps K in the unknown function. The $Poisson(5)$ prior starts out believing that there will be a good number of steps in the data. This is clearly reflected in the histograms in Figure 6.2.a for $n = 64$ and $n = 128$. The posterior mean in these cases has more "wiggle" than for the $Geometric(\frac{1}{2})$ prior even though both estimates see the same data. For $n = 512$ both estimates find that the posterior mean is roughly constant, but somewhat lower than $\frac{1}{2}$. Apparently, this is a real feature of the data and not an artifact of either prior. Eventually, for large n all of these minor considerations wash out and clear preference for very flat models has triumphed. Interestingly, from $n = 512$ or so onwards, looking at the histogram of K, the *Geometric* prior has become convinced that the model has only one split. The *Poisson* prior is only beginning to reach this level of certainty about the truth as n reaches 8192.

6.3 Theory versus Practice

Nevertheless, if the user insists upon using an estimation procedure that has been proven to be consistent (say, by the preceding theorem), he or she still have a great deal of freedom. Roughly speaking, they can use any hierarchy prior they like on the first 100!!! natural numbers and append to it an arbitrary *Poisson* tail. It is hard to imagine that there would be any practical difference between the estimators resulting from this "extended" prior and those resulting from the unmodified one, at least for realistic sample sizes. There certainly would not be any difference in practice, because in practice we only approximately compute the posterior anyway and no ordinary Markov chain Monte Carlo would ever be run long enough to notice the change.

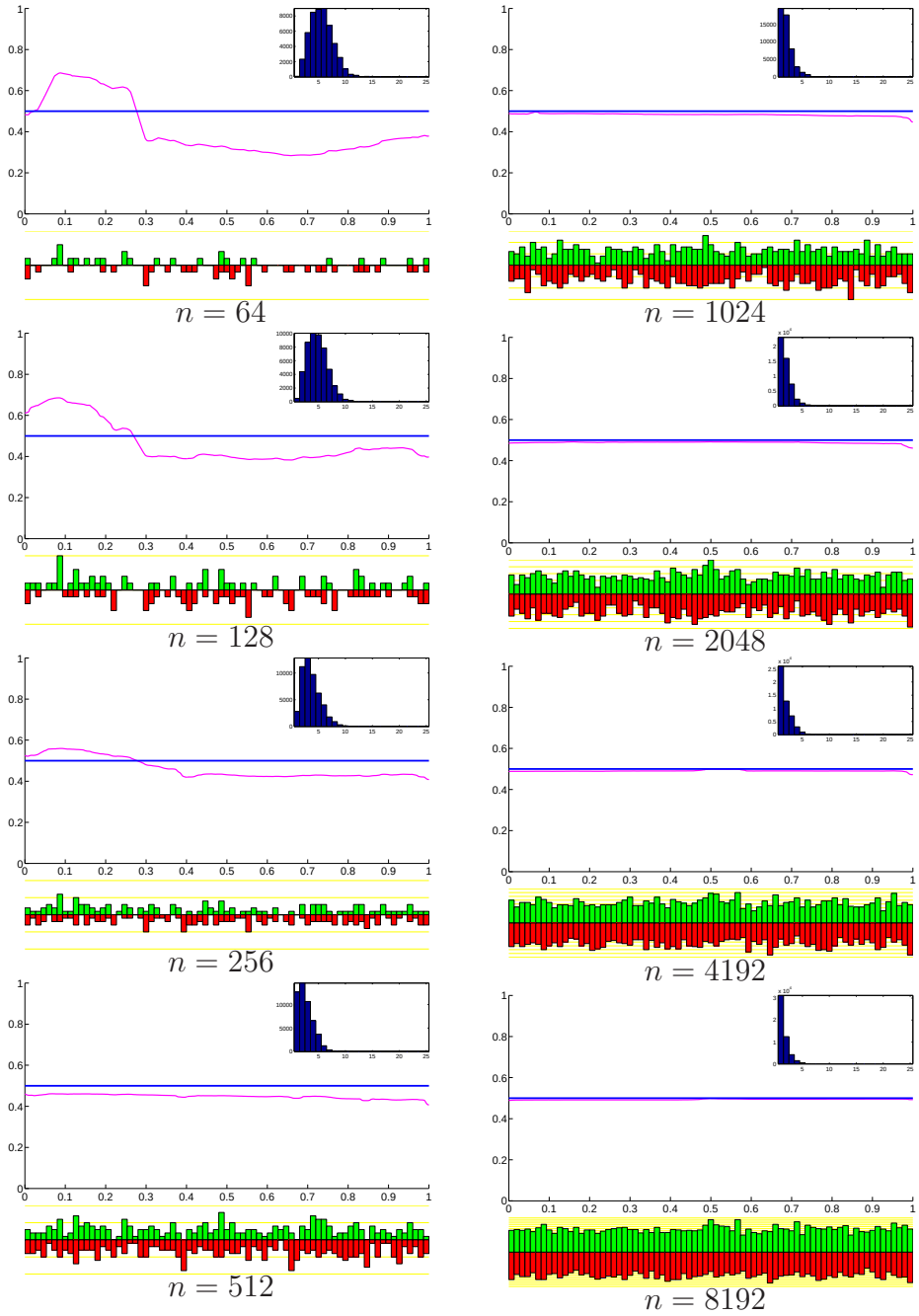

Figure 6.2.a: Null Case with $Poisson(5)$ Prior

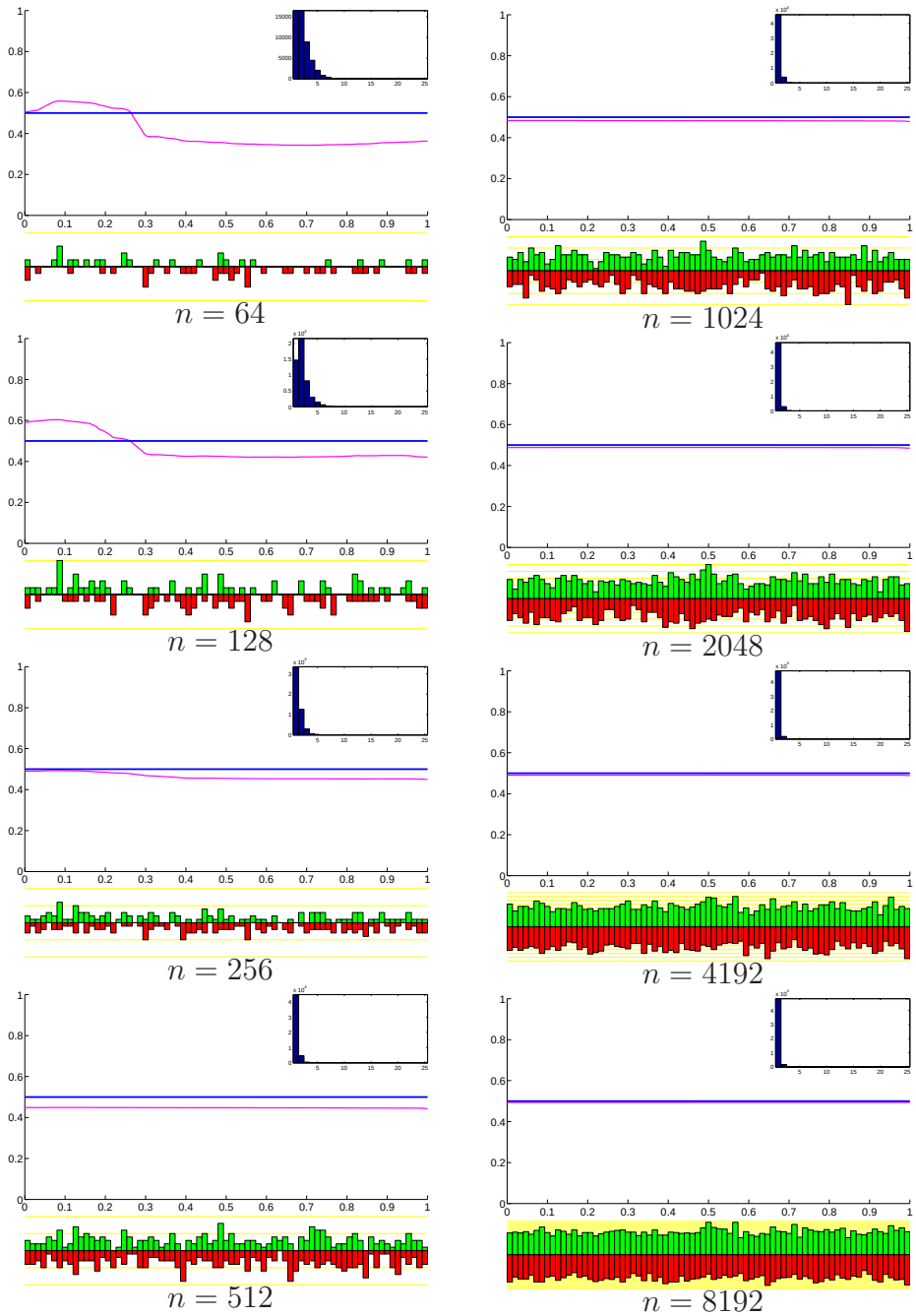

Figure 6.2.b: Null Case with *Geometric*($\frac{1}{2}$) prior

6.4 Heuristics about Poisson and Geometric Priors

Despite these concerns, I still (tentatively) advocate the use of a *Geometric* prior because of the following heuristic arguments: (a) it favors simple models (b) its tail does not drop off "too quickly" so that it will hopefully not require enormous amounts of evidence for the data to "overwhelm" the prior: by ruling out simple models in favor of better fitting models (c) its tail does drop steadily; hopefully this will protect us from "over-fitting" (d) if we consider the mode of the posterior (on a log scale), the geometric prior penalizes each additional split by a constant ($\log(\alpha)$) so that for the mode to shift to a model with an additional split we require a commensurate improvement in the log-likelihood of the data.

If using a $Poisson(\lambda)$ prior, I would be concerned that I might have specified a parameter λ that was too small. If the true regression function were unexpectedly complicated, it might take a large amount of data to overwhelm the prior. Interestingly, if we attempt to remedy this by taking an exponential mixture of *Poisson*'s, we get back a *Geometric* prior.

Another objection of mine is that (for modestly large λ) the *Poisson* prior puts less mass on models with one region than on models with two regions. This makes sense if I actually expect that the regression function will be fairly complicated, but it violates my (frequentist) training to consider a complex model before "eliminating" the simpler one.

As a compromise I would propose a prior on K that was *Geometric* until a certain point k_0 and then decays like a *Poisson*. On the other hand, if, in fact, *Geometric* priors prove reliable or even conservative, it might make sense to consider an even heavier tailed distribution. A modest proposal is to take a uniform mixture of $Geometric(p)$ priors for p ranging from 0 to 1. This results in a prior whose tail decays like $1/k$, the mass at $K = k$ being $[k(k+1)]^{-1}$.

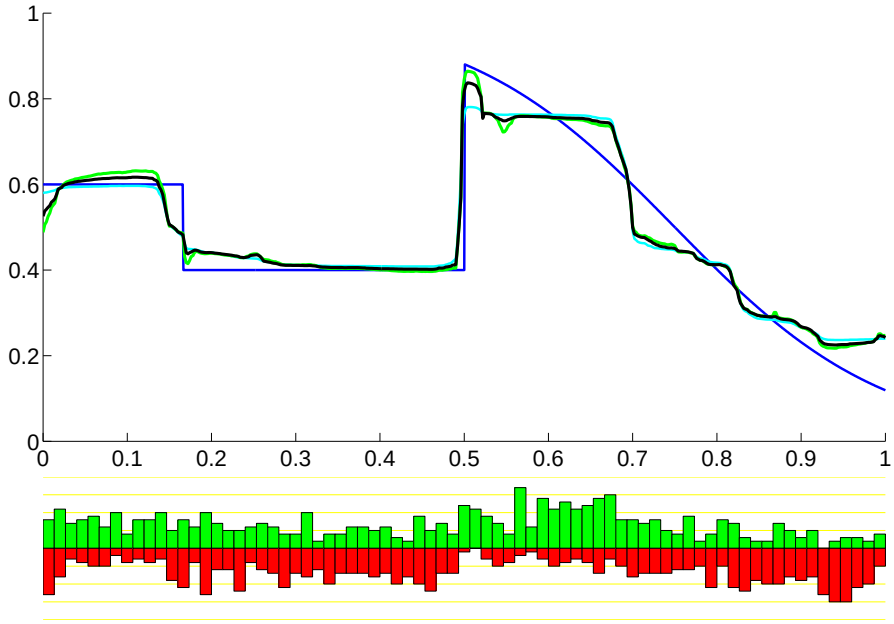

Figure 6.5.a: *The Result of Using a Poisson Prior:* Consider again the data from experimental run 1, but apply a $Poisson(\lambda)$ prior with $\lambda = 1, 5$, or 10 Key: True f (blue), $Poisson(1)$ prior (cyan), $Poisson(5)$ prior (black), $Poisson(10)$ prior (green)

6.5 Conclusions

None of these arguments is conclusive. In the absence of sound theoretical arguments it is perhaps best to rely on experimental evidence. From chapter 4 there is a good bit of evidence that *Geometric* priors perform well. What do *Poisson* priors do when applied to these data sets? In Figure 6.5.a the posterior mean resulting from a *Poisson* prior for three values of λ is plotted. The results are comparable to what happened as α was varied in Figure 4.4.a. The prior with the shorter tail flattens out some bumps that the prior with the longer tail leaves in.

Still, for some applications especially, the practical question remains: how to choose α (respectively λ)? Experience in the problem domain is the only method I can readily propose. Alternatives, like using cross-validation or an

empirical Bayes approach remain attractive, but unproven.

Chapter 7

Extensions

There are numerous ways in which to extend π, but perhaps the most pressing is to extend π to multi-dimensional data sets $D_n = \{(X_i, B_i)\}_{i=1}^n$ where the predictor X_i is in $\mathbf{R}^d$. One route to extend π to higher dimensional problems is to observe that basically, all we need to consider is a suitable way to partition the space randomly. If an interesting way to choose a partition at random is found, then describe a new prior by saying: draw a partition and give each region an independent uniform success probability. One natural way to randomly partition $\mathbf{R}^d$ is to suppose that a certain number of generating points are drawn from a Poisson process with constant rate function λ and to associate each point with its Voronoi (nearest neighbor) region. Alternatively, one could select a subset of the observed x-values at random and use their locations to determine a Voronoi partition. This alternative is, unfortunately, not a purely Bayesian proposal since the partitioning depends on the data set given. On the other hand, it only depends on the x-values of the data set, so that it remains a Bayesian procedure with respect to the response data (the y-values).

To be specific, the prior I consider (call it π^*) can be described by the following. Let $x_1, \ldots, x_n$ denote the n observed values of the covariates in $\mathcal{X}$ and let ρ be a metric on $\mathcal{X}$. Proper choice of ρ is essential to good performance in complex applications, but using Euclidean distance should suffice for simple

problems. To any subset $x_{i_1}, \ldots, x_{i_k}$ of the full list associate the Voronoi partition of $\mathcal{X}$. That is, say that a point x is in the x_{i_j} cell if $\rho(x, x_{i_j}) \leq \rho(x, x_{i'_j})$ for $j' \in \{1, \ldots, k\}$. For definiteness, in the case of ties say that x is in the cell occurring first in the original ordering of the x's. Consequently, every point $x \in \mathcal{X}$ is in exactly one cell. Put a prior on these partitions of the space $\mathcal{X}$ by putting a prior on the (finite) set of all possible (nonempty) subsets of the list $x_1, \ldots, x_n$. Say that the prior probability of a subset only depends on the size of the subset and that the probability of a given size k, is proportional to the probability that a $Geometric(1 - \alpha)$ random variable takes the value k. Finally, having chosen a subset at random, and consequently having fixed a partition of $\mathcal{X}$, assign to each element of the partition a success probability s_i drawn uniformly at random from $[0, 1]$. The generates a multi-dimensional regression function $f : \mathcal{X} \mapsto [0, 1]$ at random.

To explore these ideas I simulated a two dimensional data set with 250 data points, illustrated in Figure 7.0.a. The x-values were chosen randomly (albeit not uniformly) from the illustrated rectangle; the y-values were drawn as independent *Bernoulli* random variables whose success probability is indicated by the gray-scale in the figure. Points in the whiter regions have a higher chance of being an "x," (i.e. $y = 1$) while points in the darker regions have a higher chance of being an "o" (i.e. $y = 0$). Jointly, the x and y data was actually generated in my simulation in the reverse manner: a fair coin was flipped to determine if y will be 1 or 0 and then (conditionally) an x-value was chosen. Suppose y came up as 1, then with probability $1/2$, x will be drawn uniformly from the square on the right; otherwise, with probability $1/2$ it will be drawn from a bivariate normal distribution with standard deviation 0.1 that is centered on the left-hand square. If y came up a 0, the situation for x would be reversed. These two descriptions are essentially equivalent and the goal for the posterior mean estimate is always the same: to estimate the conditional probability of y to be 1 given x; i.e. to estimate the gray-scale image.

Sampling from the posterior of this prior is (theoretically at least) quite

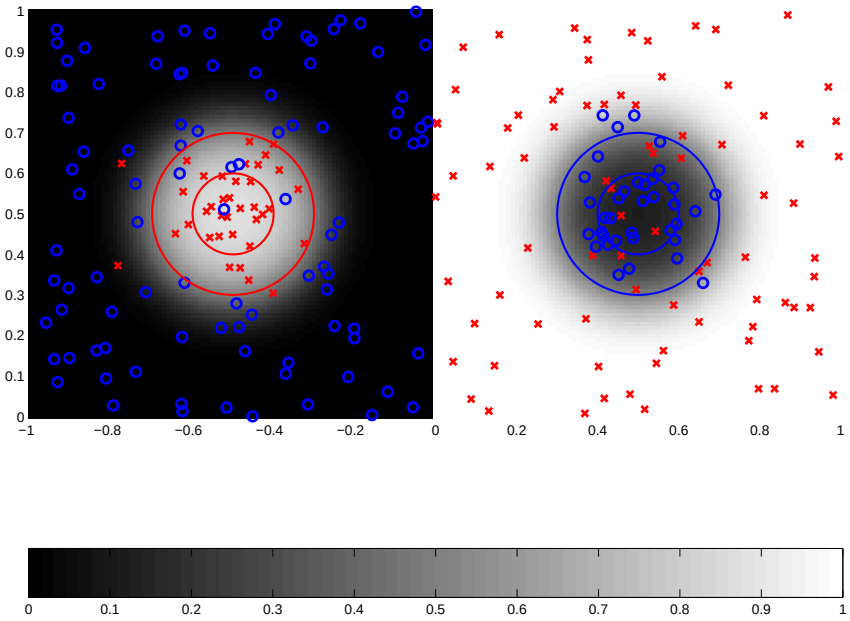

Figure 7.0.a: *A Two Dimensional Data Set and Target Function:* The gray-scale on this figure depicts a certain regression function f_0^* on a rectangle. 250 data points are drawn by flipping an $f_0^*(x)$-coin at the points indicated. Heads (red) tend to result when f_0^* is large (white). Tails (blue) tend to result when f_0^* is small (black)

simple. As before in chapter 3, the success probabilities can be integrated out analytically so that the posterior probability of a particular (nonempty) subset of size k is proportional to:

$$\alpha^{k-1} \prod_{j=1}^{k} \frac{N_j^1!\ N_j^0!}{(N_j^1 + N_j^0 + 1)!} \tag{7.1}$$

Where N_j^1 and N_j^0 denote the number of y-values equal to 1 or 0, respectively, on partition element j. Conditionally, the posterior distribution of a certain success probability $\mathbf{S}_j$ is $Beta(N_j^1 + 1, N_j^0 + 1)$. Since the number of possible subsets is finite and all of them (except the empty subset) have positive posterior probability, a standard Metropolis-Hastings type MCMC allows us to sample from the posterior (at least in theory). In practice, though, the rate of mixing matters; in an effort to improve this I have conducted preliminary work that employs the simulated tempering technique developed by Geyer and Thompson [39].

All that remains to be specified is a transitive random-walk on (nonempty) subsets of a set of n elements which has a known stationary distribution. This is easily done. Identify the class of all subsets of a set of size n with the class of binary vector of length n, with each coordinate indicating the presence or absence of a given element. Exclude the 0-vector from this set. Consider the random walk that picks a number J randomly from 1 to n and then proposes flipping the J'th bit. The proposal is not allowed if it would create the 0-vector; hold in this case. This Markov chain is easily seen to sample uniformly from the class of all non-empty subsets and consequently it is easy to modify it with the Metropolis-Hastings ratio in order to sample from the posterior.

I have also proposed extensions of this technique to put a prior on smooth functions; under this proposal the prior concentrates on functions which softly partition the space using weighted Voronoi regions, but provide for smooth

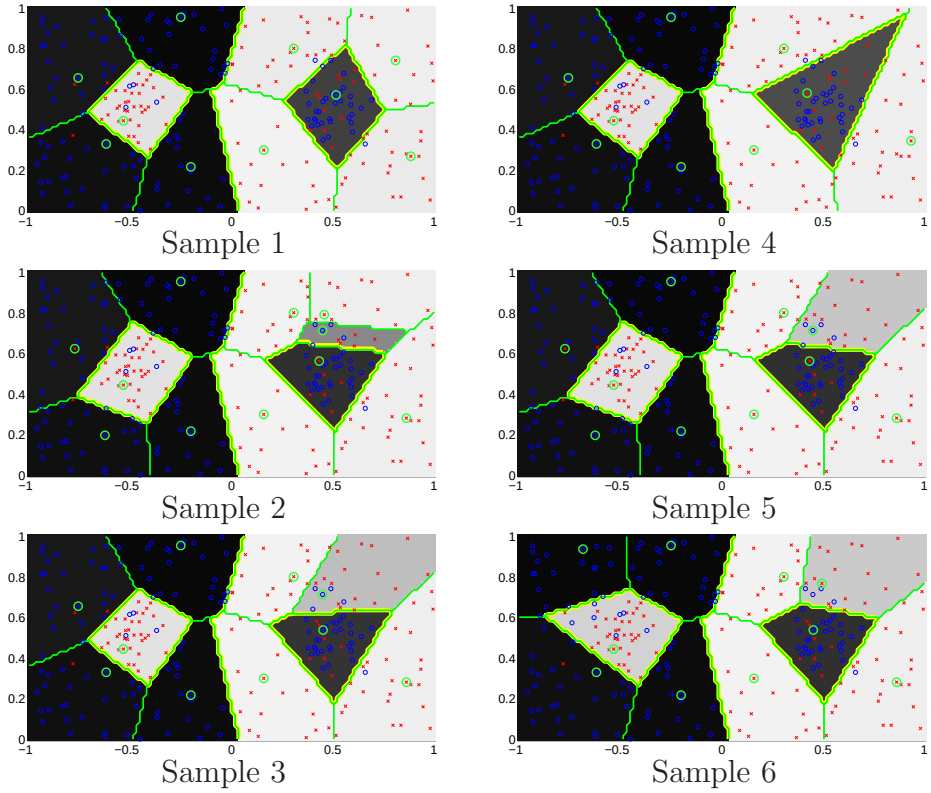

Figure 7.0.b: *Modal Samples: Voronoi Posterior* The pictured samples were the ones most frequently occurring in a sample drawn from the Voronoi posterior π^*. Each sample is equivalent to a certain subset of the original covariate list. The included points determine the partition and are drawn with a green circle. The gray-scale on a given partition element represents the posterior mean of the corresponding success probability parameter. Notice how parsimoniously the circled points determine regions that isolate out the two clumps of data.

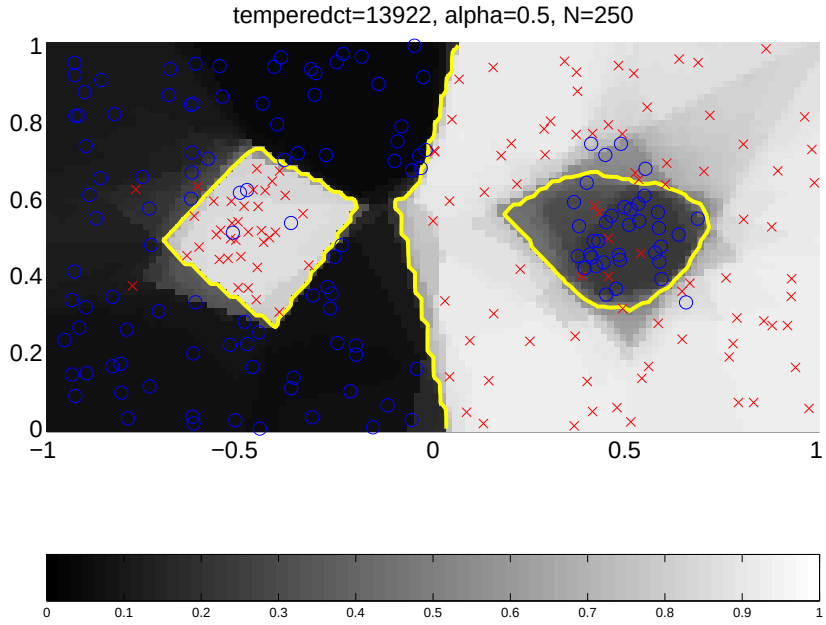

Figure 7.0.c: *A Bivariate Example.* 250 red and blue markers were put down randomly as shown. The posterior-mean estimate of f_0 under π^*, i.e. the estimated conditional probability of red at each position, is shown in gray; notice how the modal samples from Figure 7.0.b are incorporated into the posterior.

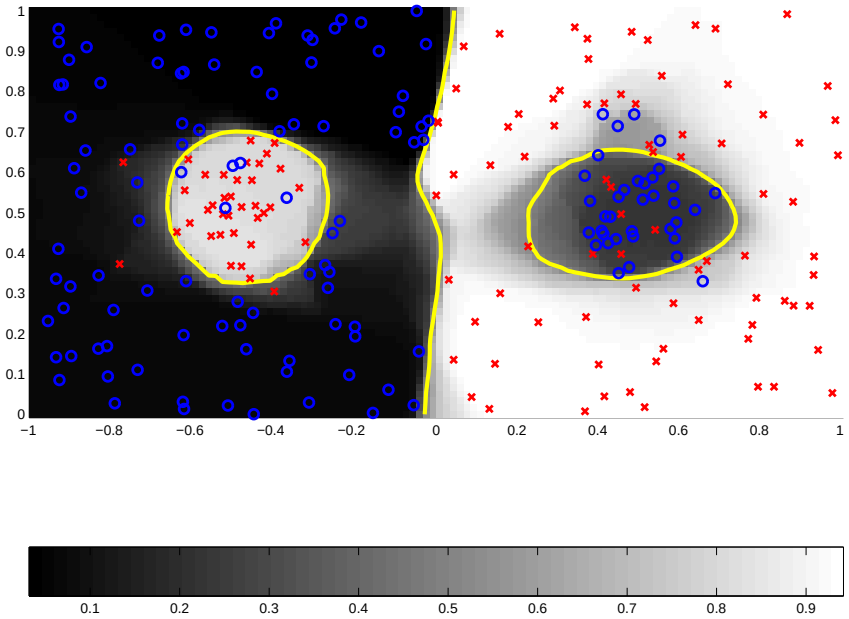

Figure 7.0.d: *Weighted Voronoi Posterior:* Redefining the Voronoi cells to use random weights allows for elliptical arcs and lines to be used in the partition and eliminates some artifacts.

transitions between regions.

A simple extension of this technique that removes many of the artifacts that are otherwise present because of the dependence on the particular locations of the covariates can be made by using a randomly weighted Voronoi partitions [54]. To make such a prior, simply augment each cell with a random variable W_j which is an independent $\Gamma(\gamma, \gamma^{-1})$ *a priori*. Then, redefine the Voronoi partition so that cells with higher weight tend to be bigger. Specifically, consider the point x to be in cell j rather than cell k whenever: $\rho(x, x_j)/w_j \leq \rho(x, x_k)/w_k$. Intuitively, if an ordinary Voronoi partition can be understood by supposing that "crystals" grow out radially from the generating seeds until they hit other growing crystals, then this weighted Voronoi partition allows for the crystals to grow at different rates. It is also simple to allow the crystals to start growing at different times (by simply adding a different random offset to ρ for each cell). The overall scale of the prior on W is irrelevant so this parameterization sets the mean to 1. For the experiment shown in Figure 7.0.d I used $\gamma = 5$. As can be seen, this simple modification allows the posterior to choose neat balls and lines with which to isolate out the different contours of the data. The posterior is computed using standard MCMC techniques.

Another route to extend π is to utilize the observed connection between π and π_{DF} (for details, please refer to subsection 2.2.1). To make that connection, I employed binary random variables $\eta_i(u)$ that indicated if a test point u was or was not above a certain random threshold V_i. To generalize, then, we can take $\eta_i(x)$ to indicate if x is in a certain random half-space H_i. This would be similar to a version of CART in which we split first into two halves via H_1, and then split each half via H_2, and so on. We could make something close to ordinary CART if we utilized random coordinate aligned half-spaces and if we employed a suitably "regularized" π_{DF} which did not assign an independent uniform to all possible 2^k binary k-tuples at level k. More generally, one can invent other ways of "regularizing" π_{DF} so that numerous binary tuples are

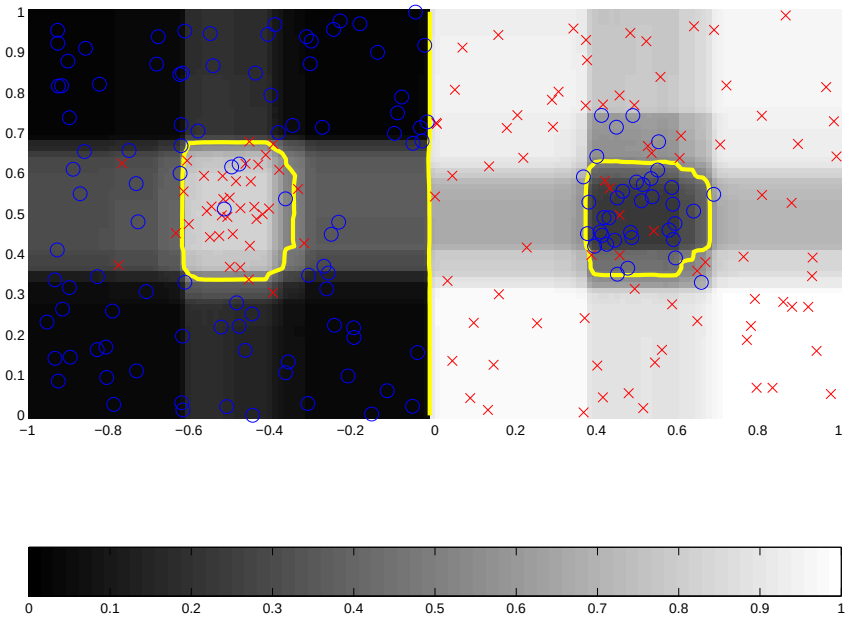

Figure 7.0.e: *Bagged CART in 2d:* Running bagged CART on this two dimensional data yields a reasonably good estimate with an interesting horizontal and vertical blurring pattern

tied together.

Finally, for comparison, Figure 7.0.e the result of bagging ordinary two-dimensional CART on this same data is shown. It has a clear advantage on the vertical split down the center that happens to be coordinate aligned. It does a decent job of isolating the two clumps of data. Interestingly, there is a clear horizontal and vertical blurring pattern that arises from the use of partitions that only partially isolated the clumped data: it is isolated in one coordinate, but not in the other.

Chapter 8

Afterword

As statisticians, we analyze data, formulate models, estimate parameters, and use these models to form predictions about future data. Broadly speaking, then, our business is inference. We go to a lot of trouble to formulate good models and we have spent a great deal of effort debating about the details of how to make inference within a model – e.g. Bayesian versus frequentist inference.

Generally, though, the way we select our model remains an art. "Non-parametric" methods are a step forward here, because, generally speaking, they at least prescribe how to select a variety of "smoothing-parameters" which essentially determine which model among some class of models we actually apply. The main topic of this thesis is of this sort: A Bayesian approach to the question of how to estimate the number and location of change-points or, more generally, how to choose a partition of the data into approximately exchangeable subsets.

In this afterword, I would like to step back from the details of this subject and address the more general problem of how we *formulate* a model. Sometimes, we prefer to shunt responsibility for this and appeal to the scientist for help; but, fundamentally, all inference comes back to data eventually – how else did the scientists discover their model? Considering, then, that the

formulation of a model is essentially "an art," and that we know that the results of our analyses are not absolute, but *relative* to the modeling choices we have made, how can we be so bold as to expect that reality will conform to our model-specific "confidence intervals?" This is not to say that statistical practice does not work, or that the formal properties of our analyses under our stated assumptions are invalid, but simply to remind the reader that the thread that connects the prescribed model with reality has no *formal* basis.

In practice, a good rule of thumb is to consider several models and, consciously or not, select one which is simple and which we expect will fit the data at least nearly as well as other more complicated models that we might prefer not to have to consider. In practice, we look at the data *ourselves* before selecting a model and build in any particular types of regularity that we happen to *notice* that it possesses into our model – at least if we think it will affect our conclusions.

Bayesian analysts do not escape these problems. They may subjectively allow for a mixture of several different models and for a range of "hyper-parameters" but this only ameliorates the core problem because no-one ever accounts for all the possible regularities that might be found.

To further make my point, consider the following, admittedly fanciful, thought experiment. An alien race comes to earth and challenges mankind to an intelligence test. We are given a binary time-series to analyze one-hundred bits at a time. It begins innocuously enough:

```
01000001000000100000110111101110000000100101000011
10101010000000001110010010000000001011001101101010
```

What statistical models shall we consider? We could try *iid* Bernoulli, or perhaps a hidden Markov model. If ambitious, we might let the data choose the order of our model. Upon doing so, we find that the null model fits best, despite the superficial appearance of runs, and we model the data as random coins with success probability 0.35. The aliens ask us to give a confidence interval on the

fraction of 1's in the next 100 bits and to estimate the probability that the 50'th bit of this new data will be 1. Most statisticians fall back on the classics and use something approximately like 0.35 ± 0.1 for both answers. Others who used a richer model suggest wider intervals and the ordinary debate ensues.

So far so good, right? Or should we worry that our glance at the data and default choice of classical model may miss structure that we failed to notice? Naah! The aliens couldn't be that tricky.... Then the next 100 bits arrive.

```
0000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000
```

A dramatic failure, but no problem, the advocates of the HMM model were ready for this sort of thing. We have, they argue, simply encountered a hidden state which always produces 0. This explains the data well enough, but some remain skeptical. They propose that the character of the data may change with every new segment of 100 bits so that the effective size of our data set is only 2. The next 100 bits arrive.

```
0001110001110001110001110001110001110001110001110001100
0111000111000111000111000111000111000111000111000111001
```

And again we are surprised. Some HMM advocates insist we just need to add a number of new hidden states. Others extend the HMM model to favor this sort of cycle-like behavior. The next 100 bits arrive.

```
0010010000111111011010101000100010000101101000110 0
0010001101001100010011000110011000101000101110000 0
```

And most are satisfied that the data has returned to *iid*, but with the world attention that this situation has generated someone notices that these are, in fact, the first 100 bits in the expansion for the fractional part of π. (C.f. `http://www.algonet.se/~eliasb/pi/binpi.html`) Oh dear – we certainly

hadn't planned on this – but we ignore this regularity in the data at our peril. It couldn't happen by chance, could it?

As the test continues, we continue to be surprised by the patterns we are sent. By now, we have learned a lot: we know less than we think about the future. Every 100 bits we have been presented with a pattern that we hadn't expected. Finally, we are given sequence after sequence that we can't explain. Eventually the aliens conclude that, however feebly, we are, at least, a modestly intelligent form of life; and, taking pity on us, they decide to reveal the patterns we hadn't discovered. The last sequences, for example, were actually Shakespeare, encoded by simple alien cipher that no human had ever considered. Furthermore, the first sequence wasn't actually "random:" to generate it, all we had to do was start `matlab` or an equivalent (alien) computational program and type:

```
x=rand(1,100)<=0.397;
sprintf('%c',x+'0')
```

Perhaps you object to my example. You prefer regression to time-series analysis and are content to consider data for which no-one would object to the model that the responses are independent given the predictors. Perhaps the problem of extrapolating a non-stationary time-series seems far too lofty to you. But you haven't escaped it by wishing it away; in fact, the time-series problem can be embedded in the regression problem. We need only suppose that the covariates are tested one by one in some fixed designed fashion so that we see the responses sequentially. If we know that the regression function "smoothly" depends on the covariates, present methods can be expected to work; but, if the dependence is sufficiently complicated, each new data point tells us something entirely new, just like in my fictitious time-series.

Even if the data is generated *iid* – the regression case considered in most of my theoretical work – there is plenty of room for improvement. In this situation, we are, indeed, much better off – we can make rigorous probability

statements about the quality of our predictions *on average*. Even so, as more and more data come in, and the general shape of the regression function becomes more tightly resolved, the "knowledge we gain" itself comes to us in a time-series fashion. Because of this, we cannot easily make predictions about the regression function at some *fixed* point.

For example, suppose the unit interval were divided into subintervals of size $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, etc... and suppose that the regression function takes a different value on each piece. In this way, we would quickly have enough data to estimate large-scale features, like the value of the regression function on the larger pieces, but if we are asked to make a prediction on one of the very small pieces, what are we to do? Perhaps, if we were smart enough, we wouldn't blithely approximate the regression function as "smooth," we would look for patterns in the regression values on the large pieces to help extrapolate to the smaller pieces. If, in fact, the regression values were seen to alternate between high and low, we would be silly to treat the function as "smooth." Instead, let us hope that were are lucky enough that it is quite "regular."

What, then are we to do? Consider again the alien's test, and the complex sequence of modeling decisions that we needed to make along the way. How can we summarize the thought process that we went through as surprising data continued to come in? What possible prior on models could our analysis (even approximately) conform to? Our only recourse, it seems, is to formalize the idea of regularity and make explicit the manner in which we choose a model to accord with the regularities apparent in the data.

A reasonable defining property of regularity is that a distribution is *regular* if it can be (approximately) reproduced within a certain budget of time by applying some modestly short computer program to the data that we have previously seen and a "random sequence." Roughly speaking, then, we can put a prior on models and/or regularities by putting a suitable prior on computer programs. Alternatively, we can select among the computer programs in a manner conforming more closely to the "method of maximum likelihood."

Formal versions of these ideas have been proposed [27],[63],[5],[47],[46],[13],[60] though much work is needed to formulate methods that are ready for actual use. Still, it seems to this author that the further development of some version of these ideas is an essential, natural, and unavoidable step in the progression of statistical thinking.

Bibliography

[1] Athreya, K. B., Doss, H. and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *The Annals of Statistics* **24** 69–100.

[2] Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics* **27** 536–561.

[3] Breiman, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140. URL `citeseer.nj.nec.com/breiman94bagging.html`

[4] Brieman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

[5] Chaitin, G. J. (1966). On the length of programs for computing binary sequences. *J. Assoc. Comp. Mach.* **13** 547–569.

[6] Chipman, H., George, E. I. and McCulloch, R. (1998a). Bayesian CART model search (with discussion). *J. Amer. Statist. Assoc.* **93** 935–960.

[7] Chipman, H., George, E. I. and McCulloch, R. (1998b). Making sense of a forest of trees. Tech. rep., Department of Statistics and Actuarial Science, University of Waterloo.

[8] Chipman, H., George, E. I. and McCulloch, R. (2000a). Bayesian treed models. Tech. rep., Department of Statistics, University of Waterloo.

[9] Chipman, H., George, E. I. and McCulloch, R. (2000b). Hierarchical priors for Bayesian CART shrinkage. *Statist. Comp.* **10** 17–24.

[10] Cleveland, W. S. and Loader, C. (1996). Smoothing by local regression: Principles and methods (disc: P80-127). In *Statistical Theory and Computational Aspects of Smoothing. Proceedings of the COMPSTAT '94 Satellite Meeting*.

[11] Coram, M. (2001). Projection pursuit for classification: A different motivation for svms. Available upon request.

[12] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *Proc. IEEE Trans. Inform. Theory* **IT-11** 21–27.

[13] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

[14] Daubechies, I., Guskov, I., Schroder, P. and Sweldens, W. (1999). Wavelets on irregular point sets. *Phil. Trans. Royal Soc. Lond. A* **257** 2397–2413.

[15] Denison, D., Adams, N., Holmes, C. and Hand, D. (2002). Bayesian partition modelling. *Comp. Statist. Data Anal.* **38** 475–485.

[16] Denison, D., Holmes, C., Mallick, B. and Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley.

[17] Denison, D., Mallick, B. and Smith, A. (1998a). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. B* **60** 333–350.

[18] Denison, D., Mallick, B. and Smith, A. (1998b). A Bayesian CART algorithm. *Biometrika* **85** 363–377.

[19] Denison, D., Mallick, B. and Smith, A. (1998c). Bayesian MARS. *Statist. Comp.* **8** 337–346.

[20] Denison, D. G. T. (1997). *Simulation based Bayesian nonparametric regression methods.* Ph.D. thesis, Department of Mathematics, Imperial College, London.

[21] Devroye, L., Gyorfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag.

[22] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation.* Springer.

[23] Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (c/r: P26-67). *The Annals of Statistics* **14** 1–26.

[24] Diaconis, P. and Freedman, D. A. (1993). Nonparametric binary regression: A Bayesian approach. *The Annals of Statistics* **21** 2108–2137.

[25] Diaconis, P. and Freedman, D. A. (1995). Nonparametric binary regression with random covariates. *Probability and Mathematical Statistics* **15** 243–273.

[26] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

[27] Donoho, D. L. (2002). The Kolmogorov sampler. Tech. rep., Stanford University.

[28] Doob, J. (1953). *Stochastic Processes.* Wiley.

[29] Doob, J. L. (1949). Application of the theory of martingales. Colloque International Ctrenat. Rech. Sci, Paris.

[30] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2002). Least angle regression. Tech. rep., Stanford University.

[31] Freedman, D. and Diaconis, P. (1984). On inconsistent bayes estimates in the discrete case. *The Annals of Statistics* **11** 1109–1118.

[32] Freedman, D. A. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. *Annals of Mathematical Statistics* **34** 1386–1403.

[33] Freedman, D. A. (1999). On the bernstein-von mises theorem with infinite-dimensional parameters. *Annals of Statistics* **27** 1119–1140.

[34] Freund, Y. and Schapire, R. (1996). *Machine Learning: Proceedings of the Thirteenth International Conference*, chap. Experiments with a new boosting algorithm. Morgan Kauffman.

[35] Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19**.

[36] Friedman, J. (1996). On bias, variance, 0/1-loss, and the curse-of-dimensionality. Tech. rep., Stanford University.

[37] Geyer, C. (1999). *Stochastic Geometry: Likelihood and Computation*, chap. Likelihood Inference for Spatial Point Processes. Wiley, 79–140.

[38] Geyer, C. and Moller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics* **21** 359–373.

[39] Geyer, C. and Thompson, E. (1993). Annealing marko chain monte carlo with applications to ancestral inference. Tech. rep., University of Minnesota, School of Statistics.

[40] Ghosal, S., Ghosh, J. and van der Vaart, A. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500–531.

[41] Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* **82** 711–732.

[42] Hasser, N. B. and Sullivan, J. A. (1971). *Real Analysis*. Dover.

[43] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.

[44] Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57** 97 – 109.

[45] Kendall, W. and Møller, J. (2000). Perfect simulation using dominating processes on ordered spaces, with application to stable point processes. *Adv. Appl. Prob.* **32** 844–865.

[46] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission* **1** 4–7.

[47] Kolmogorov, A. N. (1968). Logical basis for information theory and probability theory. *IEEE Trans. Inform. Theory* **IT-14** 662–664.

[48] Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *The Annals of Statistics* **20** 1222–1235.

[49] LeCam, L. M. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer.

[50] Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag.

[51] Metropolis, n., Rosenbluth, A., Rosenbluth, M., Teller, M. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

[52] Møller, J. (1999). *Stochastic Geometry: Likelihood and Computation*, chap. Markov Chain Monte Carlo and Spation Point Processes. Wiley, 141–172.

[53] Møller, J. and Skare, Ø. (2001). Bayesian image analysis with coloured voronoi tessellations and a view to applications in reservoir modelling.
URL `citeseer.nj.nec.com/405679.html`

[54] Okabe, A., Boots, B. and Sugihara, K. (1992). *Spatial Tesselations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons.

[55] Paddock, S. (1999). *Randomized Polya Trees: Bayesian Nonparametrics for Multivariate Data Analysis*. Ph.D. thesis, Duke University.

[56] Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

[57] Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics* **9** 130–134.

[58] Scargle, J. (1998). Studies in astronomical time series analysis. v. bayesian blocks, a new method to analyze structure in photon counting data. *The Annals of Statistics* **504** 405.

[59] Scargle, J. et al. (2002). Private communication.

[60] Schmidhuber, J. (2000). Algorithmic theories of everything. Tech. rep., IDSIA-20-00.

[61] Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **4** 10–26.

[62] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics* **29**.

[63] Solomonoff, R. J. (1964). A formal theory of inductive inference. *Inform. Contr.* **7** 1–22, 224–254.

[64] Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics* **22** 1701–1762.

[65] Vapnik, V. (1996). *The Nature of Statistical Learning Theory*. Springer-Verlag.

[66] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

[67] Wahba, G., Li, Y. and Zhang, H. (2000). *Advances in Large Margin Classifiers*, chap. GACV for Support Vector Machines. MIT Press.

[68] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve Mles. *The Annals of Statistics* **23** 339–362.